DELFT UNIVERSITY OF TECHNOLOGY

Assessment of machine learning algorithms for the purpose of heat pump detection based on load profiles and temperature readings

by

Roberto Francica

A thesis submitted in partial fulfilment for the degree
of Master of Science in Sustainable Energy Technology

supervised by:
Dr Simon Tindemans
and
Werner van Westering

July 2019

Declaration of Authorship

I, Roberto Francica, declare that this thesis titled, 'Assessment of machine learning algorithms for the purpose of heat pump detection based on voltage profiles and state estimation models' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: RFr

Date: 12th July 2019

Executive summary

The aim of this research thesis is to use machine learning models to distinguish owners of heat pumps from non-owners of heat pumps based on load profiles and temperature data only. As is the case with data mining project, its workflow can be divided into business understanding, data gathering, analysis, modeling and interpretation and deployment. As of the time of the time of this writing the models have not yet been deployed. The necessity to conduct this master thesis arises from the growing popularity of heat pumps in the Netherlands, and the potential issues associated with this spread on management of low-voltage distribution grids, in particular the rising electricity demand in the heating season. Before such issues can be analyzed however, the number of all heat pump users needs to be determined. This master thesis aims in determining precisely the overall number of heat pumps users by examining individual load profiles.

Data available for the purpose of thesis consists of load profiles of owners and non-owners of heat pumps provided by Alliander, load profiles of London-based non-owners of heat pumps, referred to as baseload profiles, load profiles of heat pumps only spread across the UK, temperature records for De Bilt in the Netherlands, London and Nottingham. The above-mentioned data was cleaned and manipulated before features were extracted from it. In particular, synthetic load profiles of heat pump owners were created by pairing baseload profiles with the pump only load profiles. Next all load profiles were normalized in order to diminish the importance of confounding variables, more on that later, and only night-hours were kept so as not to account for PV production. Such normalized load profiles were paired up in two sets: Alliander's set which consisted of load profiles of heat pump owners and non-owners provided by Alliander and simulated set which consisted of baseload profiles and synthetic load profiles of heat pump owners. It is worth mentioning that a confounding variable was present in Alliander's set, mainly the size of house since owners of heat pumps all lived in single standing houses as compared to non-owners, majority of whom lived in apartments.

Subsequently, the following four features were extracted from normalized night-time load profiles within each two sets: (1) average daily electricity consumption in January and December (this period is also referred to as winter or heating period), (2) ratio of average daily electricity consumption in January and December to average daily electricity consumption in July and August (it is also referred to as summer or cooling period), (3) slope of the curve representing mean daily temperature on x-axis and daily electricity consumption in y-axis, and lastly (4) coefficient of determination of curve

representing mean daily temperature on x-axis and daily electricity consumption in y-axis.

Three main evaluation criteria were set for the performance of machine learning models: True Negative Rate, True Positive Rate and Precision. For simplicity, the mean score was used as well, which is equal to the average of True Positive Rate, True Negative Rate and Precision. Benchmark for all evaluation metrics was set to 90%. Five models that were used to distinguish heat pump owners from non-owners were Logistic regression, Decision Tree and Support Vector Machines with Linear, Polynomial and Radial kernels. The evaluation procedure was the following: first hyperparameters for all the five models were tuned by using 10-fold cross validation with test and training set being features extracted from Alliander's set only. Next, the models with optimal hyperparameters were trained on features extracted from Alliander's set and tested on features from simulated set.

The results show that none of the models managed to reach the benchmark of triple 90% for True Positive Rate, True Negative Rate and Precision. In the hyperparameter tuning stage both True Negative Rate and True Negative Rate were close to reaching 90%, however, this has been achieved at the cost of low Precision, reaching just above 50%. This was the case due to the propensity of the models to commit type I error, that is false positives. On the other hand, at the evaluation stage when the simulated set-features served as test set, it was noticed that precision was at a significantly higher level, approximately 75%, which came at a cost of lower True Positive Rate, around 50%. True Negative Rate though did exceed 90%. These results show a strong tendency of making type II error, that is false negatives. The best performing model, which was the Support Vector Machine with Radial Kernel, achieved a mean score of 75%.

The divergence of results from the hyperparameter tuning stage to the evaluation stage is caused by the fact that there are different usage patterns of heat pumps between the owners of heat pumps in Alliander's load profiles and in the synthetic load profiles of heat pump owners. Particularly it is the case that owners of heat pumps in Alliander's set, do use heat pump in the night, as compared to synthetic users, which do so to a much smaller extent. As a result the features extracted from the simulated set of load profiles are less indicative of heat pump ownership than the features extracted from Alliander's load profiles.

This master thesis could be improved by trying out more machine learning models, improving the process of normalization of load profiles and acquiring better heat pump only load profiles which are more similar to Alliander's set in terms of usage patterns among many others. Further work can be built upon the results of this thesis. Once heat pump owners have been identified based on load profiles, similar work can be done

for identification based on voltage profiles. The advantage of using voltage profiles rather than load profiles is the fact that voltage profiles are not as privacy sensitive. Furthermore, providing more insight into the kind of heat pumps the users are utilizing (air-to-air or water source or geothermal) might provide further insight. Last but not least, the models developed in this master thesis could be deployed in Alliander and used to investigate heat pump ownership among entire dataset of 120k load profiles at the disposal of Alliander.

Acknowledgements

I would like to thank Dr Simon Tindemans of TU Delft, who supervised me throughout the thesis. His unique perspective and expertise in the field of Artificial Intelligence coupled with his knowledge in the topic of electricity grids provided me with feedback on the topics to be pursued during the thesis, and inspired me for further work.

I would like to thank Werner van Westering for being my supervisor at the company and helping me in feeling truly like home during the entire period of the internship and thesis. His guidance in terms of the company structure, and introducing me to different members of the team allowed me to gain experience in various fields within the department by broadening my horizon and uncovering the exciting world of electric grid management.

Thirdly, my sincere thanks are also directed towards Jeroen Siebers. I have had the pleasure of collaborating with Jeroen strongly during the first 6 months of the thesis work. He introduced me to Strategische Analyses team in Alliander and gave me lots of tips and suggestions as I started working on the thesis. I am sure I would not be able to jump start this research endevour as fast as I did if it was not for Jeroen.

Furthermore, I have gathered additional information regarding the use of machine learning and simulation techniques as well as data analysis with the help of the following individuals: Mathijs Danes, Barbera Droste, Gijsbert van der Geer, Jacco Heres, Robin Klaassen, Frank Kreuwel, Tom Valckx, Sidoeri Dekker, Rolf Boink, Sander Rieken, Eric Verboon and Arjen Zondervan.

Contents

Acronyms

AUC Area Under Receiver Operating Characteristics. 25, 28, 29, 61, 64, 65, 67, 68

CRISP-DM Cross-industry standard process for data mining. 7, 8

DNO Distribution Network Operator. 2, 40, 84

DT Decision Tree. 14–19, 31, 64, 65, 71, 72

FNR False Negative Rate. 24, 27

FPR False Positive Rate. 24, 26, 28

IALM Intrusive Appliance Load Monitoring. 10, 11

KDD Knowledge Discovery in Database. 7, 8

LR Logistic Regression. 14–16, 31, 63–65, 71, 72

NIALM Non-intrusive appliance load monitoring. 10, 11

SVM Support Vector Machine. 14, 15, 20, 23, 24, 31, 65–72, 81

TNR True Negative Rate. 24, 26, 61–65, 67–71, 74, 80, 81

TPR True Positive Rate. 24, 26–28, 61–65, 67–74, 80–82

Dedicated to my family and friends

Chapter 1

Introduction

1.1 Chapter Introduction

The goal of this chapter is to introduce the problem that the research presented in this thesis aims to tackle. Furthermore, research objectives and (sub)questions as well as methods for achieving those objectives are stated. This is followed by the outline of the thesis report.

1.2 Background

In the most basic definition of the device, a heat pump is an appliance which transfers the thermal energy from its source into heat sink. Heat pumps thus transfer the thermal energy in the opposite direction than spontaneous heat flow, by absorbing energy from cold space and transferring it into a hot one.

While the technology itself is not a new one - the first ground source heat pump was developed by Robert C. Webber in the 1940s and the first large-scale implementation dates back to the inauguration of London's Royal Festive Hall in 1951 - the wide-scale diffusion did not occur until 1970's when the price of electricity skyrocketed and consumers started to look for alternatives ways to heat their homes (1).

According to (2) in 2011, heat pump industry recorded approximately 6.3 billion dollars of revenue while in 2017 this amount increased by some 90% to the value of 12 billion dollars. This trend is bound to continue as concerns over global warming increase and several initiatives, such as the Paris Agreement (3), are being implemented in order to tackle the rising temperatures across the planet.

Netherlands, which is a signatory of the Paris Agreement, has decided to take additional steps on its own apart from the ones delineated in the document. One such step is the gradual phase-out of exploration of the Groningen gas fields by 2030. Furthermore, the government has vowed to transition all residential buildings off-gas by 2050 (4). It is important to recognise that not only does the central government take steps to lower the emissions, but so do individual municipalities like Utrecht, Amsterdam and Rotterdam by giving green light for construction of neighbourhoods completely disconnected from national gas network (4).

As mentioned in (4) "in the Netherlands 38% of energy consumption goes to heating. Half of this is used by residential buildings. And 89% of Dutch houses have a gas-fired boiler. All in all, residential heating contributes to some 10% of Dutch CO2 emissions." The task of transitioning 89% of households from gas-based heating to heat-pump is a tremendous task, and preparatory work needs to be done in order for it to succeed. Such preparatory work is conducted first and foremost by the Dutch Distribution Network Operator (DNO) - Alliander.

1.3 Alliander's role in the master thesis

I have had the pleasure of working as intern Data Scientist in Alliander from the very beginning of the thesis. I was part of the *Strategische Analyses* team which goal was to predict how the spread of different technologies - photovoltaic technologies, electric vehicles and heat pumps - will impact the network considering its limits.

The importance of this research to the Dutch DNO should be elaborated upon. As mentioned in the previous section, the prospect of massive number of Dutch households transitioning to heat-pump based heating requires preparatory work to be done before the advent of the trend. This preparatory work includes studying the potential effects of proliferation of heat pumps into the distribution grids as in (5), (6), (7) and (8). The goal of such research is to make sure that investments in distribution grid are made to accommodate the rising volume of transmitted electricity required to operate the heat pumps.

The adoption of heat pumps will not take place overnight however. Adoption rates will vary based on country, regions, cities and even neighbourhoods. It will be thus necessary to identify the current heat pump users and make the required grid adjustments in those areas in which heat pump usage is proliferating at a quick pace. The benefit of identifying such users is that the next logical step is to identify those households which have similar characteristics as the owners of heat pumps and thus are most likely to

adopt heat pump in a short time-frame. In other words by providing a clear indication of who already has heat pump, it is possible to predict how heat pump usage will spread as the time passes. This in turn allows to directly answer the question regarding upgrading and investment in electricity grid - which is the key business area for Alliander.

1.4 Problem definition

Most of the methods for detecting heat pump owners using machine learning are based on data obtained from individual household's load profiles and a pre-existing database of load profiles of singular devices such as heat pumps as detailed in (9), (10), (11) and (12). Unfortunately to perform such pattern matching task, it is necessary to posses load profiles of frequency greater than 1 Hz. The load profiles that Alliander has access to are in 15-minute intervals, corresponding to approximately 0.001 Hz, making it impossible to perform analysis similar to the ones above.

Furthermore, there is a high probability of bias being present in the data provided by Alliander. For starters, households with heat pumps tend to be larger in terms of size and number of inhabitants. Furthermore, those houses tend to be new-built thus potentially excluding fragments of populations from lower income bracket. In order to account for those confounding variables not only will Machine Learning be used to detect heat pumps on the dataset provided by Liander, but also new data will be simulated, thus lowering the probability of presence of confounding variables.

1.5 Research objectives and questions

The objective of this thesis is to provide a method of classifying each household as either an owner/user of a heat pump or not. This problem needs to be solved assuming that (1) the smart meter data is available in 15-minute intervals, i.e. extremely low frequency, and (2) the dataset originally provided by Alliander possess confounding variables which might distort the working of the machine learning model upon deployment.

In order to systematise the problem research (sub)questions have been formulated. the main research question is:

Is it possible to accurately predict presence of heat pumps based on load profiles from smart meters?

Sub questions:

1. Which classification model performs best for the purpose of heat pumps detection based on load profiles derived from smart meters?
2. Do simulated load profiles of heat pump users based on open-source datasets of heat pump load profiles and load profiles of gas-heated households possess the same characteristics as the real ones?
3. How do the machine learning models tested on synthetic load profiles compare to models trained and tested on smart-meter based load profiles?

1.6 Methods

The research objective will be achieved in three steps. In the first step labelled dataset of load profiles of both heat pump owners and those who do not own this device is obtained and features are extracted. By the end of this step several machine learning algorithms such as Support Vector Machines, Logistic Regression and Decision Tree would have been evaluated in terms of performance metrics and their optimal hyper-parameters will be chosen.

The second step revolves around simulating load profiles of heat pump owners from two separate datasets: (1) dataset of heat pump only load profiles; (2) dataset of load profiles of gas-heated households - that is not heated by a heat pump. Furthermore, once synthetic load profiles are generated, machine learning algorithms will be used to classify each of the customers as owner or not of a heat pump. Results from this step and the first step will be compared and significant divergences between the two parts will be discussed.

1.7 Outline

The following chapter, chapter 2.1, presents detailed literature review, divided into three topical segments - data mining projects' workflow, heat pump detection using machine learning algorithms and an overview of literature on simulating load profiles.

Chapter 3 delves into the theory of machine learning, evaluation procedures and metrics.

Chapter 4 provides an overview of the datasets used in the research. Considering that correlation between electricity consumption and temperature will be investigated, a

publicly available dataset with meteorological figures will be discussed. Furthermore, the two datasets used to simulate load profiles will be studied.

Chapter 5 presents an overview of the simulation process used to create synthetic load profiles of heat pump owners, as well as analysis of the simulated datasets.

Chapter 6 focuses on the choice of features and their numerical analysis.

Chapter 7 discusses the experimental procedure and the final choice of evaluation metrics. Furthermore it delves into the results of the classification models, by discussing the numerical outcomes and explaining the divergences between the outcomes achieved for Alliander's set based features and the simulated ones.

Chapter 8 presents the conclusions of the research, provides suggestions for further work and summarises the thesis.

Appendix A presents plots of every experimental procedure outlined in chapter 7, for those seeking more detailed information.

1.8 Chapter summary

This chapter summarised history of the development of heat pumps, their role in lowering of carbon dioxide emissions, alongside discussing the coming transition of the Netherlands from a gas-reliant country to completely off-gas in the decades to come.

Considering that there is a difficulty in obtaining load profiles of both users and non-users of heat pumps without including biases, this thesis aims to simulate unbiased load profiles of heat pump users such that no biases are present.

The end goal of the thesis will be achieved in two steps: (1) using machine learning algorithms with smart-meter derived load profiles coupled with temperature data to classify each client as owner or not of a heat pump. Next (2) synthetic load profiles of heat pump users will be simulated to ensure that no biases are present and machine learning will be applied to synthetic load profiles to evaluate whether the feature set and machine learning models used for the meter-based load profiles are applicable to unbiased data.

Chapter 2

Literature Review

2.1 Chapter introduction

This chapter summarises the findings of literature which is associated with the topic of this master thesis. This review is divided into four topics: structuring of data mining project, overview of the smart meter data analytics, heat pump detection and load profile simulation. This section will based on literature discussed in (13).

2.2 Addition of heat pumps to the grid

The effects of addition of heat pumps to the grid have been studied in (6). As the UK Renewable Heat Premium Payment, a UK government programme aimed at subsidising the instalment of heat pumps, was rolled out, data from the installed was collected. That data in contained electricity consumption data of heat pumps.

According to (6) four potential problems arise due to mass deployment of heat pumps:

1. Drastic increase in peak electricity demand;
2. Ramp rate increase - generators will need to increase the speed with which they will output power;
3. Voltage drop beyond allowed limits and
4. Overheating of low voltage feeder and transformer due to insufficient thermal capacity

According to the calculation performed in (6), the After Diversity Maximum Demand, which is a measure of peak demand, of a 700-strong population of heat pumps amounts to 1.7 kW. This peak occurs in the morning at around 7am. It was confirmed that there is strong relation between outside temperature and heat pump electricity consumption as shown on figure 2.1.

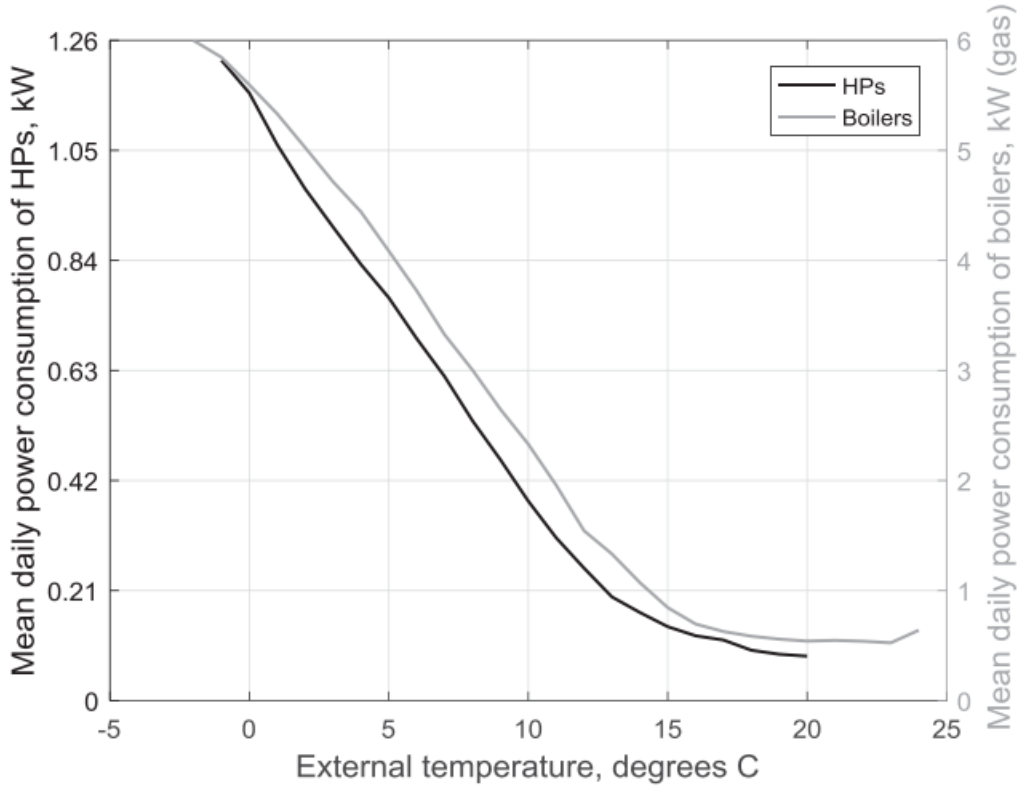

FIGURE 2.1: Curve of mean daily power consumption of heat pumps versus external temperature. Source: (6)

Overnight, heat pumps reach some 40% of its peak consumption with the second peak occurring in the evening. It was simulated that if 20% of British households were using heat pumps to heat their spaces, grid's peak demand would increase by 14%.

2.3 Workflow of a data mining project

This thesis can be described as a data mining project, or as it is known otherwise, a data-driven knowledge discovery, Knowledge Discovery in Database (KDD), project first described in (14). KDD requires a certain workflow to be used when performing related project work, as displayed on figure 2.2. The above-mentioned steps have been widely adopted in the industry and are today referred to as Cross-industry standard process for data mining (CRISP-DM) (15).

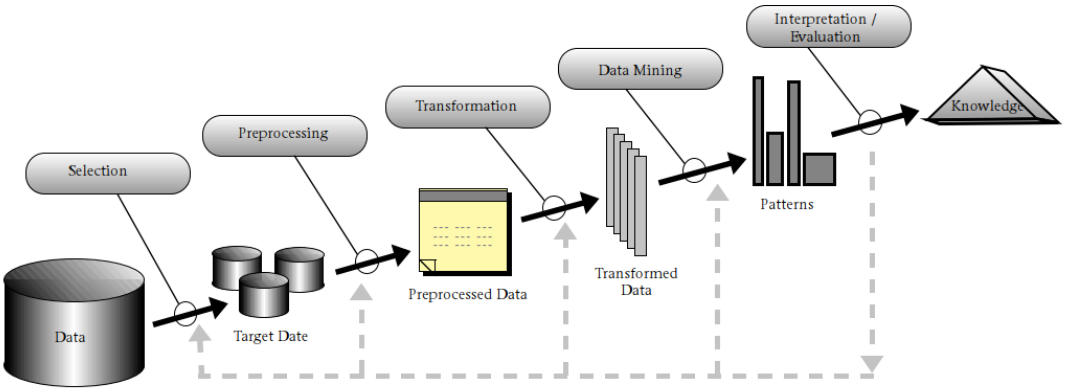

FIGURE 2.2: An Overview of the Steps That Compose the KDD Process. Source: (14)

In order to better understand the relation of KDD to this research project the following list maps the activities performed in this master thesis to the CRISP-DM steps:

1. Selection: datasets were provided by Alliander and were extracted from open-sources such as data.gov.uk;
2. Preprocessing: Data was cleaned, missing values were treated;
3. Transformation: Load profiles of heat-pump owners was simulated by coupling baseload profile and heat-pump only profiles. Furthermore, features from simulated dataset and dataset provided by Alliander are calculated;
4. Data mining: Data was analysed in order to understand which features would allow for best prediction capability in terms of distinguishing between heat pump owners and those who do not own such a device
5. Interpretation/Evaluation/(Implementation): Using machine learning models to distinguish between owners and not of heat pumps. Selecting the optimal model;
6. (Optional for research purposes, obligatory in commercial environment) Deployment: Using the best performing machine learning model chosen int he previous step in order to indicate which customers have heat pump load profiles using the entire database of Alliander's heat pump load profiles;

2.4 Smart meter data analytics

Ever since the inception of smart meters, electricity consumption data started being analysed to extract characteristics of customers, both on household and aggregate levels of feeders and transformers.

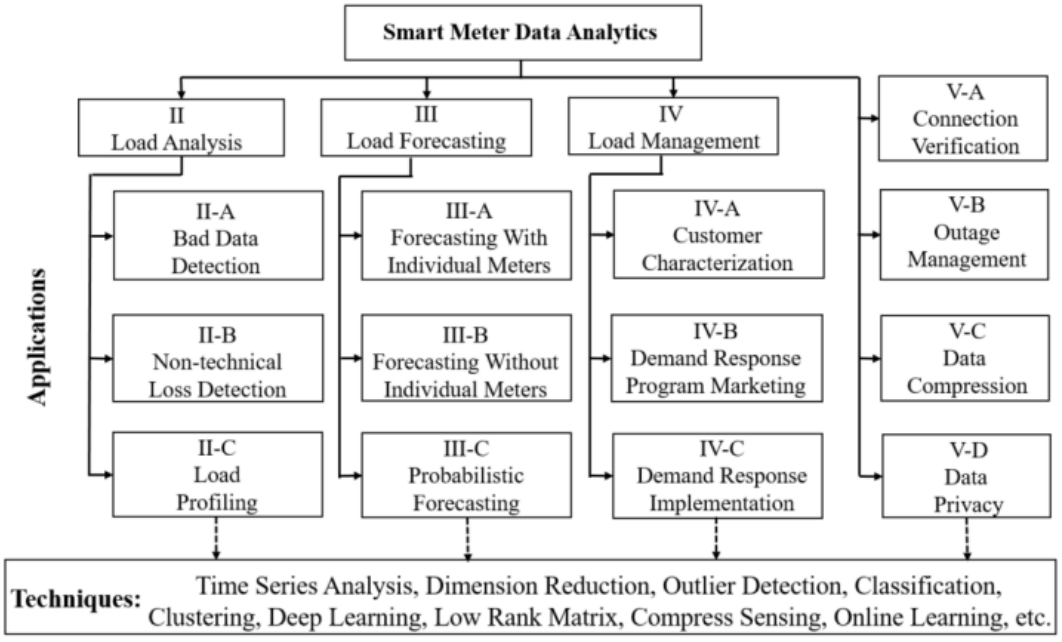

FIGURE 2.3: Taxonomy of smart meter data analytics as in (13)

As it can be seen in figure 2.3, smart meter data analytics can be divided into four broad categories:

- Load analysis,
- Load forecasting,
- Load management and
- Others

This thesis falls within the category of Load Profiling, section II-C on the 2.3 figure, that is extraction of characteristics regarding grid users from their load profiles. Goals of customer characterisation usually serve the purpose of understanding what impact electricity consumption - such as certain devices. This information is subsequently fed into the load predicting models, section III on the 2.3 figure. The logic goes that the more information is available about the users, the easier will it be to predict within acceptable degree of accuracy their future electricity consumption.

In (16) customers are classified into 16 different categories based on their frequency coefficients. (17) aimed to determine which factors and clients characteristics influence to the greatest extent electricity consumption of a given load profile: residents' personal characteristics, resident's socio-economic factors, stock and holdings of electrical appliances, household structural characteristics or residents behavioural factors. Other works, such as (18) focus solely on determining from load profiles socio-economic factors.

2.5 Heat pump detection

Within the Load Profiling branch of smart meter analytics, a subbranch called Appliance Load Monitoring has emerged in the recent years. There are several goals of appliance load monitoring, which include predicting the energy usage of a particular device based on past data or detecting a particular device from the aggregate load profile of a household or a part of it, the latter being the focus of this thesis.

Appliance Load Monitoring can be divided into two types: Intrusive Appliance Load Monitoring (IALM) and Non-intrusive appliance load monitoring (NIALM).

2.5.1 Intrusive appliance load monitoring

IALM make use of a separate electricity reader/meter placed within given household, compound, etc. to read the electricity flow. According to (19), IALM can be divided into separate three fields (each of them characterising the placement of the electricity meter):

- *1 meter for a zone*. In this case a single meter is placed to collect readings from a part of the house
- *1 meter for a plug*. In this case a single meter is placed to collect readings from a single plug-in.
- *1 meter for appliance*. A meter placed within the appliance.

It is important to note that for the purposes of device recognition IALM makes the analytics part of the recognition easier as by definition as either (I) there is access to electricity readings directly form the device of interest or (II) the electricity reader is placed at a less aggreagte level compared to a smart meter reading the electricity flow from entire household. Furthermore IALM data, especially the *1 meter for appliance*

FIGURE 2.4: Overview of the *1 meter for appliance* Intrusive appliance modelling process as in (1)

can be used to validate the NIALM environments (20) (21). IALM is also used to offer a better understanding into the user's consumption behaviours and aids in predicting the energy usage for a given appliance (22) and (23). The obvious disadvantage however is that the costs of monitoring are significantly higher due to the large number of installed sensors.

2.5.2 Non-intrusive appliance load monitoring

As mentioned in (1) the NIALM is a method "where an electrical circuit that contains a number of devices which are switched on and off independently - can be monitored." One of those devices could be a heat pump, detection of which is the goal of this thesis.

FIGURE 2.5: Overview of the NIALM process as in (1)

Research within the field of NIALM dates back to the 1980s and so far several subdivisions of this branch have appeared. These subdivisions can be differentiated by the following factors (24):

- Frequency of Measurements: frequency of measurements can range from aggregate 1-hour measurements to MHz range as in (25), (26) and (27). It is necessary to point out that there is an inherent trade-off between frequency and storage economics. The more frequent the data the more memory is required to store it. On the other hand more frequent data indicates naturally greater visibility, especially into on/off devices. In this thesis data is collected within 15 minute intervals.
- Real/Reactive Power: Some works use only real power as in (25) whereas others utilise readings of both real and reactive power as in (28). In this thesis only real power is being considered.

- Use of external features: some research does include external factors such as time of the day, month of the year or temperature, with the last factor being used in this thesis.

- Supervised/Unsupervised Training: Supervised training assumes obtaining labelled dataset indicating which load profile(s) correspond to which device(s) we are trying to detect. Unsupervised learning can be used as well in absence of labelled dataset as in (29). The third option is to use semi-supervised machine learning in which part of the dataset is positively or negatively labelled while the rest remain unlabelled as in (30). In this thesis supervised machine learning techniques, used for binary classification, will be used.

- Training/Testing Generalisation: Most previous work is focused on detecting a specific device in variety of conditions - for example detecting a given washing machine in a set of load profiles. Rarely however has it been attempted to generalise those findings to scan for different devices of the same category within various load profiles. This thesis aims to generalise across different types of heat pumps rather than a specific model.

- Evaluation Metrics: The most standard evaluation metric is the accuracy - in terms of classification problems and it has been applied in variety of research papers as in (25), (29) and (27). However, other non-standard metrics such as the percentage of energy misclassified are also used as in (31). This thesis uses the following metrics: True Negative Rate, True Positive Rate, Area Under Receiver Operating Characteristics and Precision.

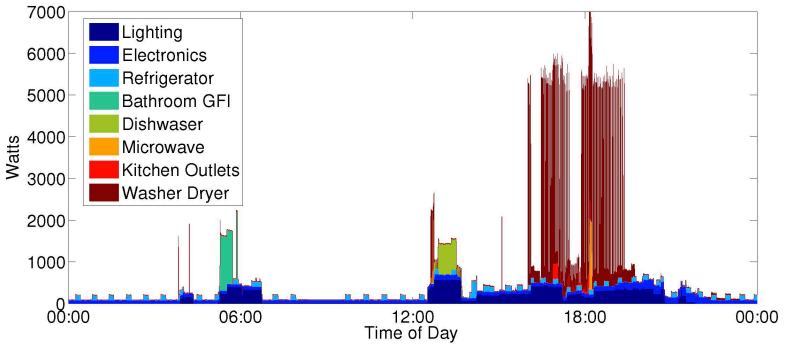

FIGURE 2.6: An example of energy consumption over the course of a day for a house as in (24)

2.6 Simulation of load profiles

Simulation of load profiles is not as widely-used of a technique in the industry considering the advent of widely available load profiles online such as (32), (33) and (34). Nevertheless some literature can be found on this topic and as it might be relevant to this thesis, it will be subsequently discussed.

The goal of simulating load profiles might be to predict based on simulated data the aggregate load profile of a neighbourhood/zone etc. as in (35) and (36). This is particularly useful method when faced with lack of individual load profiles. In both cases the input information needed to construct a typical load profile consists of shares of annual energy consumed for heating, cooling, lighting and appliance, average home size of given area, socio-cultural factors such as ethnic background, income level, etc. The model then subsequently outputs based on the input data a typical load profile within a given area.

Unfortunately neither the goal nor the methods of simulating load profiles in the above-mentioned literature is applicable to this thesis. Firstly, the goal in this thesis is to simulate not a typical load profile of a heat pump owner but a diverse set load profiles, all of which have characteristics of a load profile. Next, information on area size and socio-cultural factors are not available for the purpose of this thesis.

2.7 Chapter summary

In this chapter, literature pertaining to smart meter data analytics, heat pump detection and load profile simulation has been discussed. In particular, it was noted that smart meter data analytics can be divided into four broad categories Load analysis, Load forecasting, Load management and Other. Subsequently, Appliance Load Monitoring and its application to this thesis was elaborated on followed by brief overview of literature available on the topic of load profile simulation, used frequently for aggregate load profile characterisation.

Chapter 3

Overview of relevant data mining techniques

3.1 Chapter Introduction

In this chapter the theoretical aspects, main principles and primary applications, of machine learning will be discussed. In particular the three algorithms used in this research, Logistic Regression (LR), Decision Tree (DT) and Support Vector Machine (SVM)s will be discussed in detail. Lastly evaluation criteria for model selection and principles of cross-validation will be elaborated on.

3.2 Machine Learning

The field of machine learning can traditionally be divided into several sub fields. One of those fields is supervised machine learning. It is this area that will be discussed in detail in the following sections.

3.2.1 Theoretical principles of supervised machine learning

The idea behind supervised machine learning revolves around *learning by example*. What is meant by that is that under a hypothetical situation in which we want an algorithm to be able to detect chairs, we would supply the algorithm first with a sample of chair and NOT chair pictures, and subsequently based on those pictures such an algorithm would then learn what a chair is and thus would be used to detect one once a new image is provided.

More formally, as in (37):

For each observation of the predictor measurement(s) x_i, $i = 1, \ldots, n$ there is an associated response measurement y_i. We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference).

3.2.2 Binary classification problems - principles and evaluation criteria

There are several problems which can be solved using supervised machine learning. These problems include classification - designating an observation into a given category, and regression - predicting a numerical outcome. In this research, binary classification, meaning classification between two categories, takes place.

Some examples of common binary classification problems include:

- Spam detection - algorithms indicate whether a given email is a spam or not;
- Fraud detection - algorithms indicate whether a given transaction is suspicious or not;
- Medical testing - indicating whether a given patient has a disease/illness or not;
- Quality control - indicating whether a given product is manufactured well enough to be sent to clients or not;

3.3 Classification algorithms

There are several machine learning algorithms used for purpose of binary classification. These models differ from one another in the way those models "through iterative optimisation of an objective function [...] learn a function that can be used to predict the output associated with new inputs" (38). The three different types of machine learning algorithms used in this research - namely LR, DT and SVMs - will be discussed in the following sections.

3.3.1 Logistic Regression

LR is an algorithm which outputs the probability of occurrence of each observation in the dataset given that the observation is positive. This implies that the LR uses a

function which outputs values between 0 and 1 for all possible inputs. The function that the LR uses is called logistic function, as seen on figure 3.1 and it has the following form:

$$p(X) = 2 * \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Where X is the vector of observation variables(features), β_0 is the intercept and β_1 is the vector of feature coefficients.

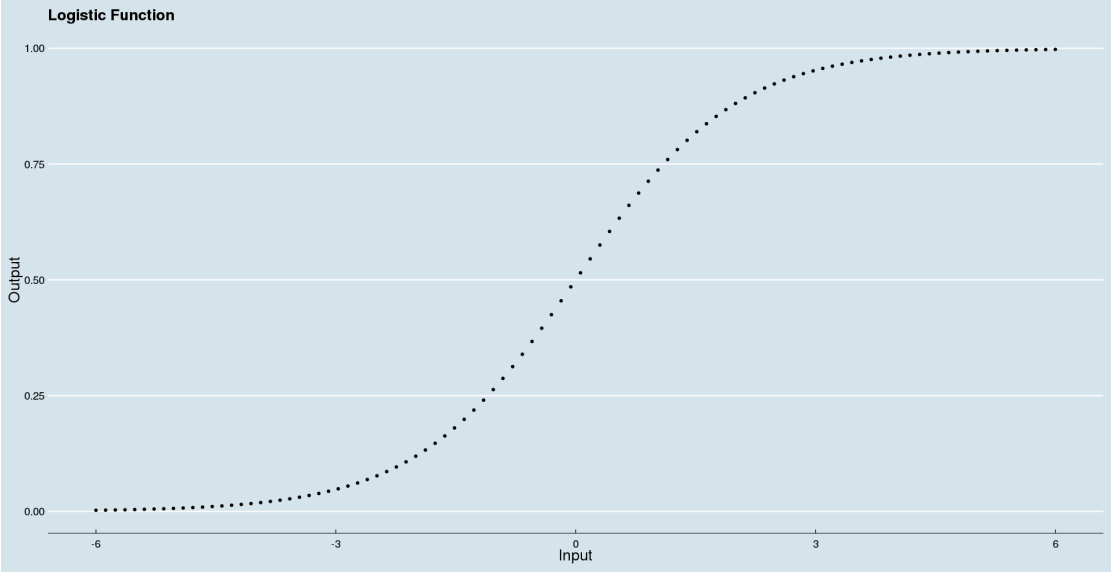

FIGURE 3.1: Logistic Function for input values ranging from -6 to 6

We can rearrange the above equation in the following way:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} -> log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X$$

We can clearly see here that there is a non-linear relation between the observation values and the odds, the ratio of probability to one minus the probability. While in a least squares regression, the algorithm chooses the β_1 and β_0 by minimising the square error, in case of LR, the algorithm chooses the β_1 and β_0 by minimising the maximum likelihood function, as displayed below:

$$Max.likelihood = \prod_{i=1:y_i=1} p(X_i) * \prod_{j=1:y_j=0} (1 - p(X_j))$$

3.3.2 Decision Trees

The idea behind DT revolves around segmentation of predictor variables into separate groups based on features. This segmentation can be easily represented via a tree, thus

the name DT. An example of a DT can be seen on 3.2. In this case we can see that there are 3 splits based on two variables - balance and income.

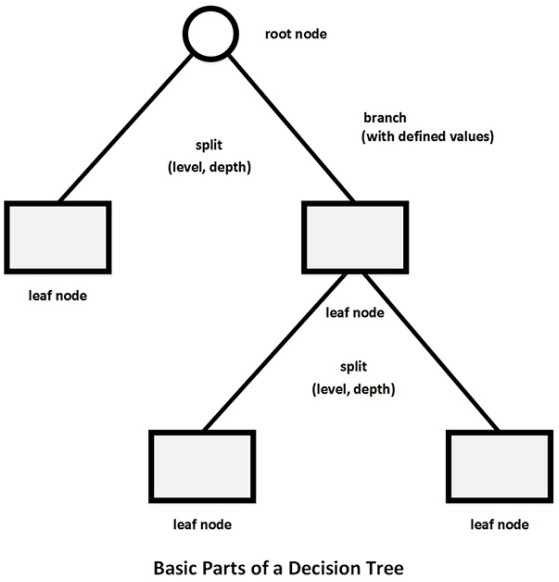

FIGURE 3.2: An example of a DT. Source: (39)

It is necessary to understand how are those splits made. In other words, why are the splits made on the basis of balance<1000 rather than balance<900 or any other numerical value? Here come into play two main DT algorithms, CART and C4.5. They differ on what criteria are used when choosing which split to be made at a given instant.

Tree splitting is done on the criteria of either information gain or harmonic mean of the Gini impurity index in given node. For example, let's assume that at a given node there are have 250 positive observations and 150 negative observation. Furthermore, it is decided to split based on the following categorical variables:

- student - with possible answers Yes (200 positive observations, 50 negative) and No (50 positive observations and 100 negative ones);
- married - with possible answers Yes (50 positive observations, 125 negative) and No (200 positive observations and 25 negative ones).

In the case above, Gini impurity index for each of the branches above will be calculated in the following way:

- For variable student:
  - For leaf node yes the Gini impurity index is: $1 - \frac{200}{200+50}^2 - \frac{50}{200+50}^2$=0.32
  - For leaf node no the Gini impurity index is: $1 - \frac{50}{100+50}^2 - \frac{100}{100+50}^2$=0.44

  - For both nodes the weighted Gini impurity index is:
    $\frac{150}{150+250} * 0.44 - \frac{250}{150+250} * 0.32$=0.365

- For variable married:

  - For leaf node yes the Gini impurity index is: $1 - \frac{50}{125+50}^2 - \frac{125}{125+50}^2$=0.41
  - For leaf node no the Gini impurity index is: $1 - \frac{200}{200+25}^2 - \frac{25}{200+25}^2$=0.20
  - For both nodes the weighted Gini impurity index is:
    $\frac{225}{225+175} * 0.20 - \frac{175}{225+175} * 0.41$=0.29

Thus it can be seen that for variable married the weighted impurity index is lower than for the student variable, thus it shall split using the married variable. The splitting using the Gini impurity index is conducted within the tree until the minimum number of cases within the lead node is reached. Frequently however, DTs are being split to such an extent that they become overfit such as in figure 3.3, even after specifying the proper minimum number of cases within a leaf.

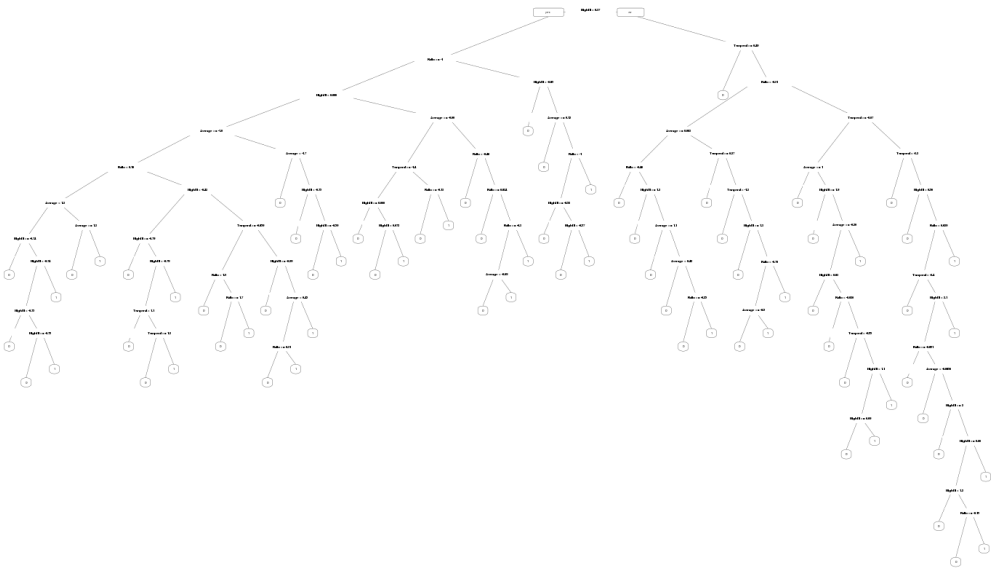

FIGURE 3.3: An example of a large unpruned tree

In order to prevent overfitting, tree pruning is conducted with the aim of reducing the size of the tree -that is the number of splits. The algorithm for tree pruning is the following:

1. Construct a DT;
2. For every single number of splits, evaluate the classification error for a 10 fold cross validation. Report the mean and standard deviation, as in figure 3.4

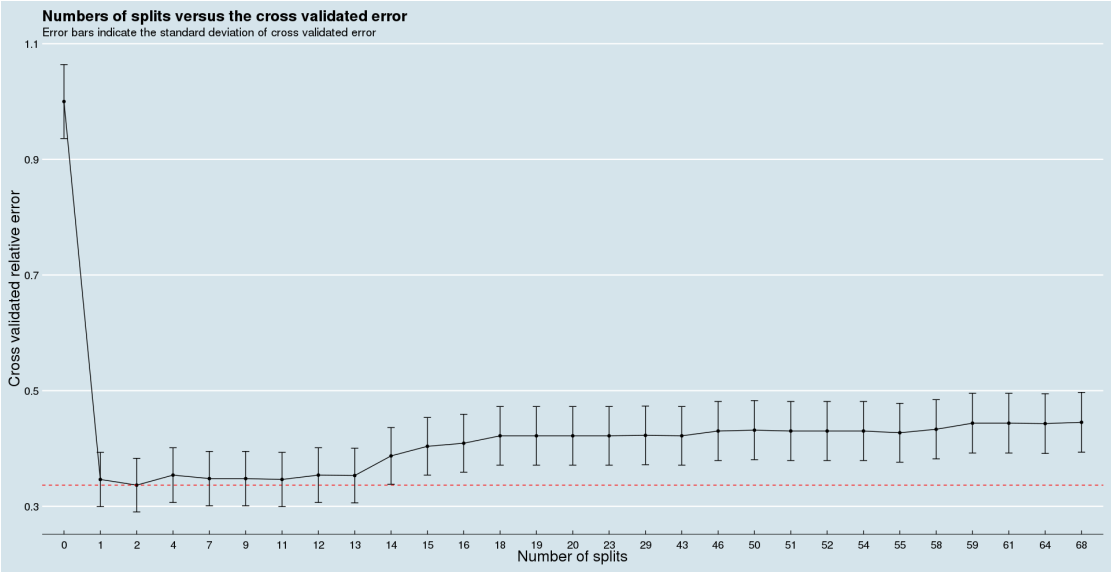

FIGURE 3.4: Plot of relative cross-validated error versus the complexity tree. Red line indicate the level of the lwoest cross-validated error

3. Chose such an amount of splits that corresponds to the lowest cross validated mean misclassifcation error.

An example of a pruned version of the DT presented in figure 3.3 can be seen on figure 3.5.

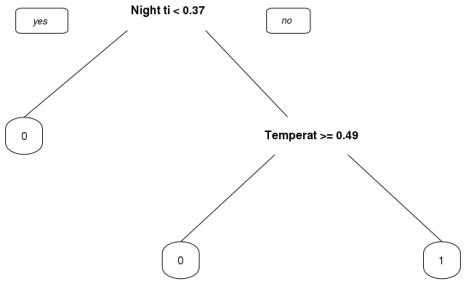

FIGURE 3.5: An example of pruned version of the tree on 3.3

3.3.3 Support Vector Machines

SVM is another example of a widely used algorithm in the industry. In order to understand its advantages and use cases its simplest form will be explored first, *Maximal margin classifier*. Subsequently *Support Vector classifier* will be explained and finally its most complex form, that is *SVM* will be defined.

3.3.3.1 Maximal margin classifier

In order to understand the notion behind the Maximal margin classifier, the concept of hyperplane needs to be elaborated first. In a p-dimensional space, a hyperplane is a flat affine subspace of dimension p-1 (37). An example of a hyperplane can be seen on figure 3.6. As it can be seen the area of the hyperplane in grey satisfies the following relation: $3X - Y + 2 < 0$ while the non coloured area satisfies the following relation: $3X - Y + 2 > 0$.

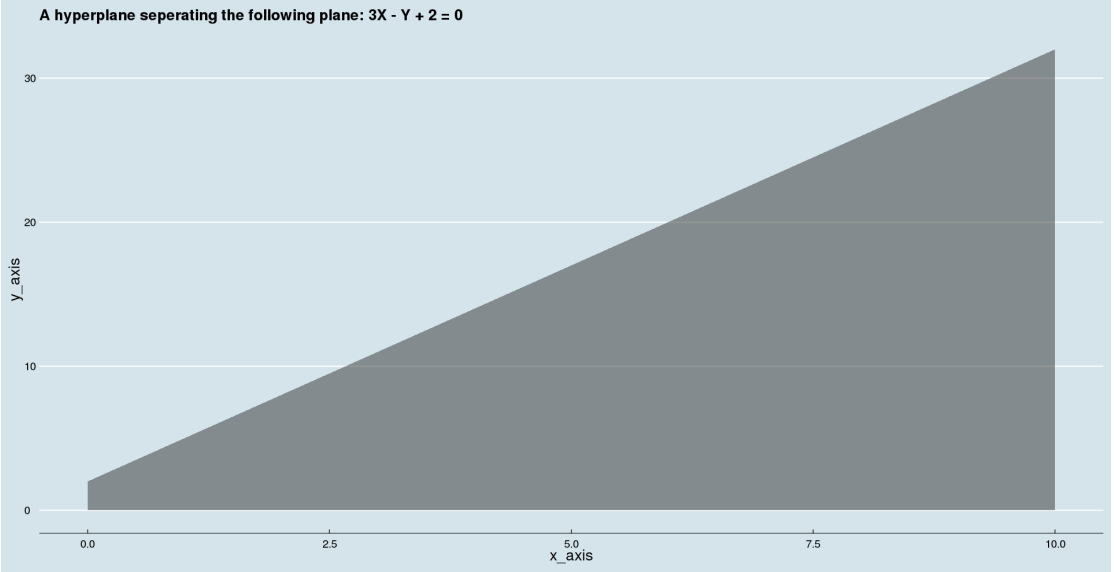

FIGURE 3.6: An example of a hyperplane

Let's imagine now that there are two classes in our dataset, both of which can be perfectly separated by using a hyperplane as in 3.7. The idea behind a maximal margin hyperplane is to find such a hyperplane boundary so as achieve maximum possible distance between the boundary and the closest point to the boundary. As it can be seen, this is an optimisation problem which can be expressed in the following way:

for a hyperplane:
$\beta_0 + X_1\beta_1 + ... + X_p\beta_p$

Maximise M
with the following optimisation variables $\beta_0, \beta_1, ..., \beta_p$
subject to:
(1) $\sum_{j=1}^{p} \beta_j^2 = 1$ and
(2) $y_i(\beta_0 + X_{1,i}\beta_1 + ... + X_{p,i}\beta_p) \geq M \vee i = 1, ..., n$
(3) $M \geq 0$

Where p is the number of variables, n is the number of observations, y_i is the actual class of the i-th observation. Equation (1) ensures that the distance between the closest point and the boundary is calculated in the perpendicular way while the equation (2) ensures that every observation is not only on the appropriate side of the boundary but also at a distance M from it - given that the margin is greater or equal than 0. An example of the Maximal Margin classifier is shown on 3.7. One feature of maximal margin classifier is that only those observations which are the closest to the boundary, at distance M, impact the classifier. If any of the points not at distance M to boundary of the Maximal Margin classifier were displaced, the solution would not change.

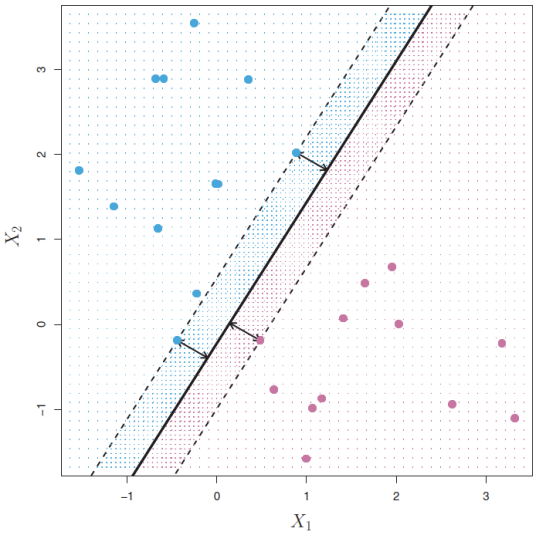

FIGURE 3.7: An example of a separable hyperplane. Source: (37)

The fundamental problem with Maximal Margin Classifier is two-fold:

- if no single hyperplane can separate the two classes, there is no solution to the optimisation problem.
- addition of a single observation at a distance lower than M, completely changes the solution of the optimisation - extreme sensitivity.

3.3.3.2 Support vector classifier

Support Vector Classifier aims to deal with the issue of non-separability and vulnerability to changes in single observations. It does so by allowing for there to be not only observations on the 'wrong' side of the margin but also on the wrong side of the boundary, as in 3.8. The purpose of allowing for the two phenomena to occur is two-fold:

- to avoid overfitting;
- to deal with non-separable cases as well.

For that purpose a new optimisation problem is designed designed which has the following form:

for a hyperplane:
$\beta_0 + X_1\beta_1 + ... + X_p\beta_p$
Maximise M
with the following optimisation variables $\beta_0, \beta_1, ..., \beta_p$
subject to:
(1) $\sum_{j=1}^{p} \beta_j^2 = 1$ and
(2) $y_i(\beta_0 + X_{1,i}\beta_1 + ... + X_{p,i}\beta_p) \geq M(1 - \epsilon_i) \vee i = 1, ..., n$
(3) $M \geq 0$,
(4) $\epsilon_i \geq 0 \vee i = 1, ..., n$
(5) $\sum_{i=1}^{n} \epsilon_i \leq C$;

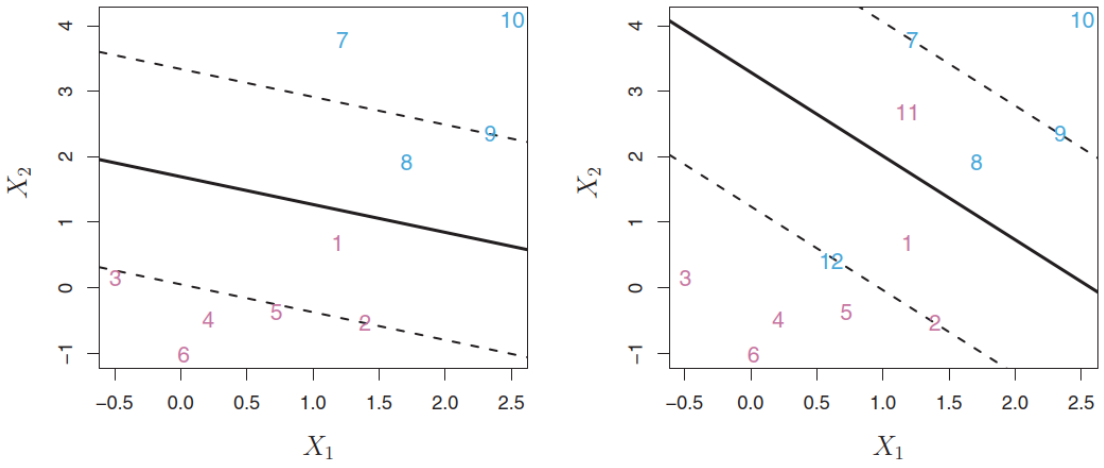

FIGURE 3.8: An example of a support Vector Classifier boundary. Source: (37)

The only changes that are present in support vector classifier and maximum margin classifier relate to the presence of ϵ known as slack variables and the tuning parameter C. The role of the former is to allow for the observations to be on the wrong side of

the margin or the hyperplane, while the latter constraints the number of violations of the margin or the boundary that can occur - this parameter is known as cost parameter, also referred to later on as simply C, and will be tuned in chapter 7. Thus it can be seen that the tuning parameter C controls the bias-variance trade-off as arguably the more violations of both the margin and the hyperplane are allowed for, the less sensitive to singular observations our classifier becomes, at the expense though of potential increase in bias, as always is the case.

In Support Vector Classifier, as the name suggests we are also dealing with support vectors, however in this case they have a different meaning. Support vectors in support vector classifier are those observations that lie directly on the margin or on the wrong side of the margin for their respective class. Thus once again it is a classifier that is impacted only by a handful ob observations rather than the entire dataset.

3.3.3.3 Support Vector Machine

So far however the boundaries are still linear. There might however be situations in which the boundaries might be nonlinear as in figure 3.9. In order to deal with nonlinear hyperplanes, the feature space is enlarged using kernels - functions of different type which enlarge those feature spaces.

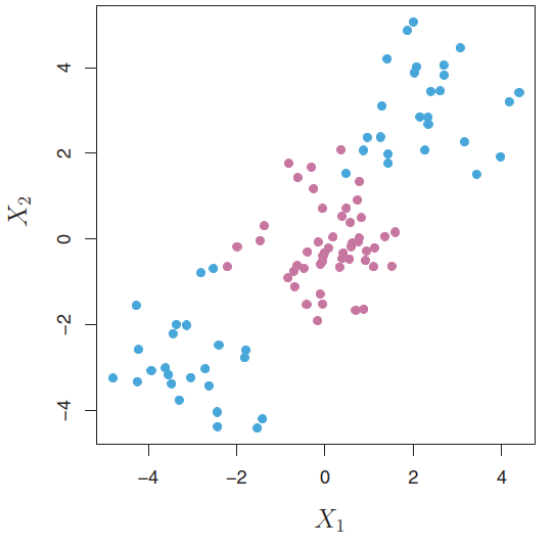

FIGURE 3.9: An example of classification problem that cannot be solved via a linear boundary. Source: (37)

Those kernels can be of polynomial form, radial, sigmoid or linear (in which case the classifier is a Support Vector Classifier). In this thesis polynomial and radial SVMs will be used and in case of the former the degree of the hyperplane will be one of the tuned hyperparameters. For the purpose of this thesis, kernels will not be explored

in detail, but it is important to understand that in those cases the kernels are used for non-linearly separable hyperplanes as in 3.10. Furthermore, the linear Support Vector Classifier will be hereafter referred to as linear SVM or SVM with linear kernel.

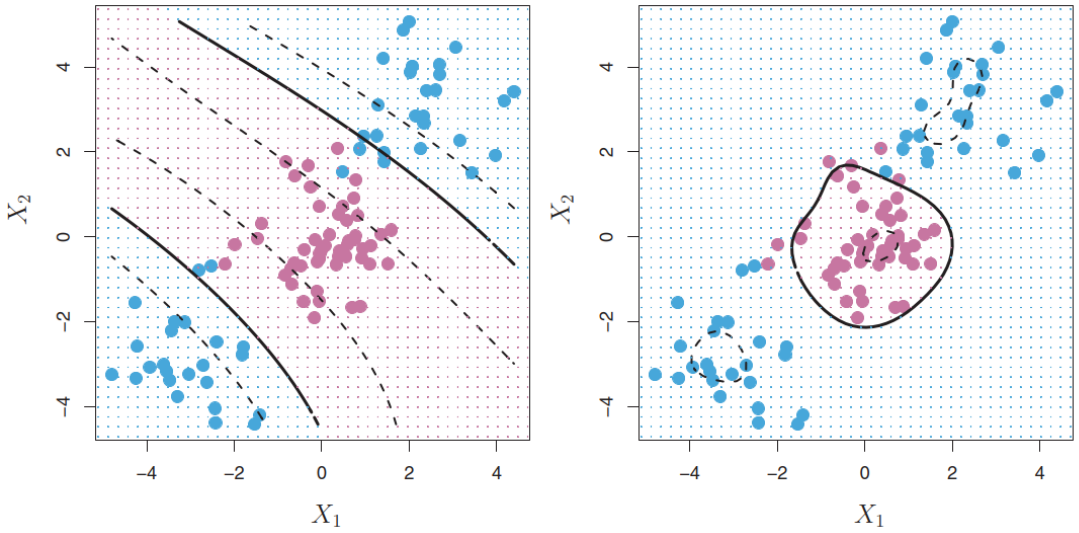

FIGURE 3.10: An example of classification problem solved via SVM with polynomial, left-hand side, and radial, right hand side, kernels. Source: (37)

3.4 Evaluation criteria

Once an algorithm has been constructed with all of its properties set, in which case it becomes a classifier or more broadly a model, its performance needs to be evaluated. The overall question that the evaluation criteria ought to answer is: *How well did the classifier perform?*

The following valuation criteria will be discussed:

1. *Accuracy*;
2. *True Positive Rate (TPR)*, also known as *Sensitivity* or *Recall*;
3. *False Positive Rate (FPR)*, also known as *fall-out*;
4. *True Negative Rate (TNR)*, also known as *Specificity* or *Selectivity*;
5. *False Negative Rate (FNR)*, also know as *Miss Rate*;
6. *Precision*
7. *F1 Score*, also known as *F-score* and *F1-score*;

8. *Area Under Receiver Operating Characteristics (AUC)*, also known as *AUROC* or Area Under ROC Curve;

Before however details of each of the evaluation criteria can be discussed, a concept known as confusion matrix, as seen on figure 3.11, has to be introduced. In binary classification one of the categories is designated as positive and the other one as negative. As it can be seen there are four possible combinations of actual values/classifier predictions for a binary classification:

- True Positive - both the actual value of the observation and the classifier value are positive;
- False Positive - the actual value of the observation is negative while the classifier value is positive;
- True Negative - both the actual value of the observation and the classifier value are negative;
- False Negative - the actual value of the observation is positive while the classifier value is negative;

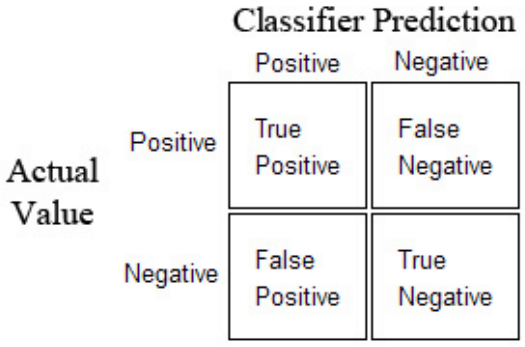

FIGURE 3.11: Example of a confusion matrix. Source: (40)

Ad 1. Accuracy

The ratio between the sum of true positive and true negative observations and the total number of observations. It assesses what percentage of all observations were correctly classified.

$$Accuracy = \frac{\#TP + \#TN}{\#Observations}$$

Ad 2. TPR

The ratio between the true positive observations and the sum of all actual positive observations. It assesses what percentage of all actual positive observations were correctly classified.

$$TPR = \frac{\#TP}{\#TP + \#FN}$$

Example: What percentage of all patients with diabetes did we classify as sick?

Ad 3. FPR

The ratio between the False positive observations and the sum of all Actual Negative Observations. It assesses what percentage of all actual negative observations were incorrectly classified.

$$FPR = \frac{\#FP}{\#FP + \#TN}$$

Example: What percentage of all patients without diabetes were classified as sick?

Ad 4. TNR

The ratio between the true negative observations and the sum of all actual negative observations. It assesses what percentage of all actual negative observations were correctly classified.

$$TNR = \frac{\#TN}{\#FP + \#TN}$$

Example: What percentage of all patients without diabetes were classified as healthy?

Ad 5. FNR

The ratio between the false negative observations and the sum of all actual positive observations. It assesses what percentage of all actual positive observations were incorrectly classified.

$$FNR = \frac{\#FN}{\#TP + \#FN}$$

Example: What percentage of all patients with diabetes were classified as sick?

Ad 6. Precision

The ratio between the true positive observations and the sum of all positive classifier predictions. It assesses what percentage of all positive classifier predictions were correctly classified.

$$Precision = \frac{\#TP}{\#FP + \#TP}$$

Example: What percentage of all patients classified as sick actually suffer from diabetes?

Ad 7. F1 Score

F1 score is a measure of trade-off between classifying correctly all positive observations and classifying incorrectly all negative observations. It is calculated by taking the harmonic average of *Precision* and *TPR*.

$$F_1 = 2 * \frac{Precision * TPR}{Precision + TPR}$$

Ad 8. Area Under Receiver Operating Characteristics

Numerous binary classification algorithms, output not only actual predictions, negative or positive, but also probabilities that a given observation is positive. As such, these probabilities have values ranging from 0 to 1. It is for those probabilities that

a Receiver Operating Characteristic Curve, with FPR on x axis and TPR on y axis, is constructed in the following way:

1. For X (X is an integer) unique equally spaced threshold values between 0 and 1:
   (a) For every observation, if the probability of the observation being positive is greater than or equal to the threshold, assign this observation as positive, otherwise as negative.
   (b) For such predictions, calculate FPR and TPR.
2. Plot a piece-wise constant plot of TPR versus FPR as seen in figure 3.12.
3. Calculate the area under the curve.

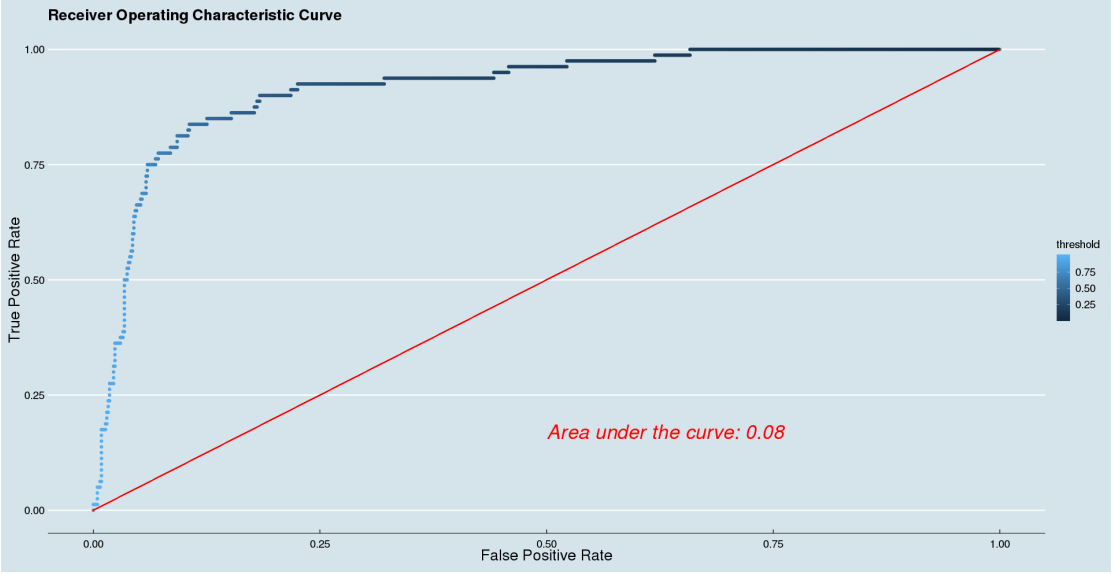

FIGURE 3.12: Example of a Receiver Operating Characteristic Curve

For threshold equal to 0, all observations are assigned as positive, thus the TPR is 1, however the FPR will also be equal to 1, as all observations are classified as positive even though there are some negative observations. As we increase our threshold, an increasing number of observations will be classified as negative causing the FPR to decrease. For threshold equal to 0, all observations are classified as negative and thus FPR and TPR are both equal to 0.

The higher the AUC the better the algorithm performs. To understand what the AUC represents, an ideal Receiver Operating Characteristic Curve is displayed in figure 3.13. In this situation it cam be seen that for thresholds other than 1, the classifier is capable of always detecting all positive cases. If AUC is equal to 0.5 - indicated with the red line on figure 3.12 - then it represents random guessing. If the area under the curve is smaller

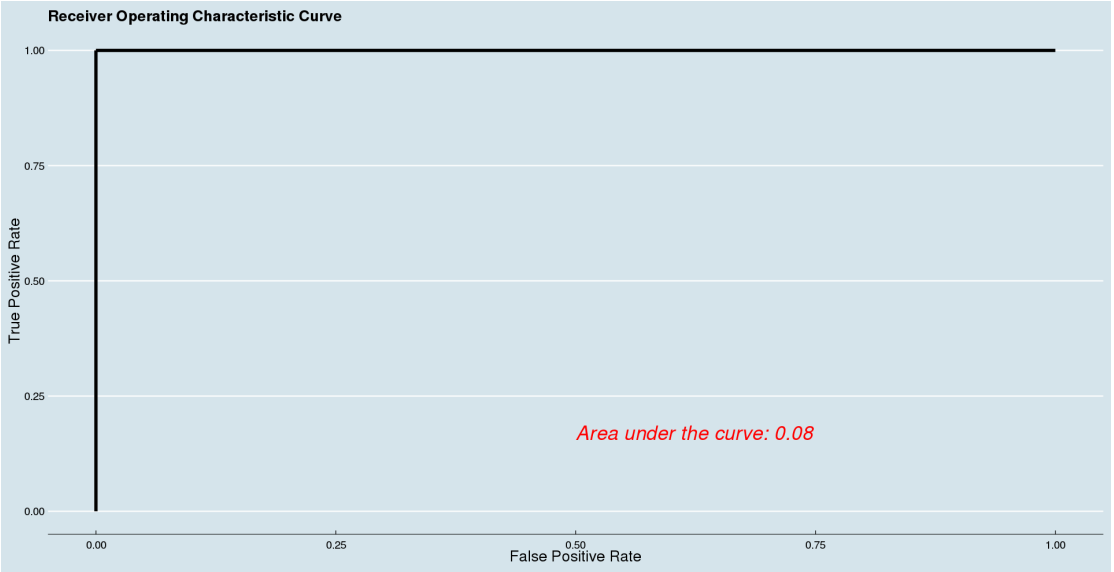

FIGURE 3.13: Example of an ideal Receiver Operating Characteristic Curve

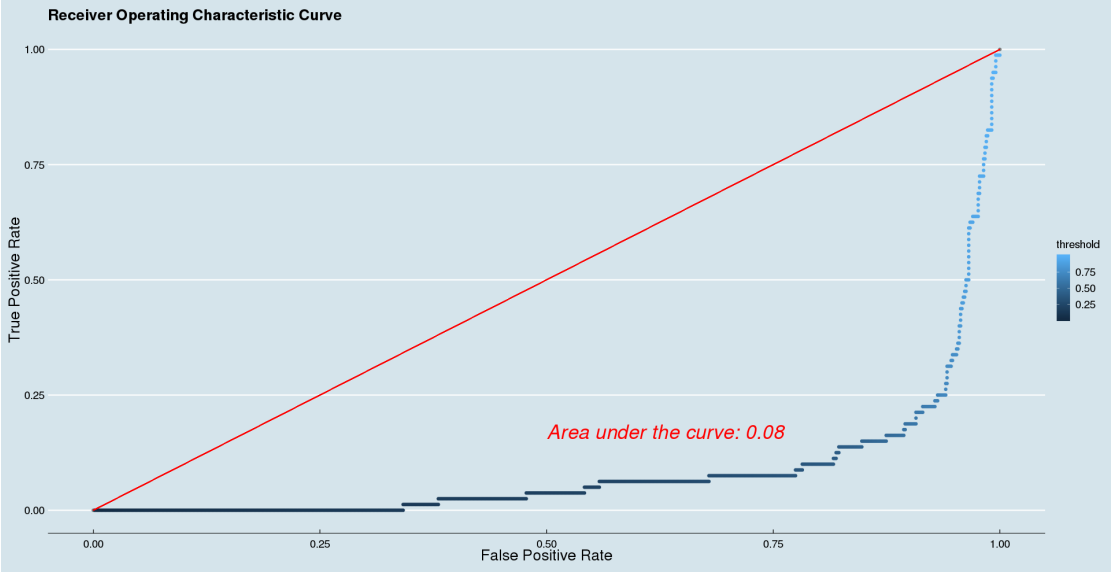

FIGURE 3.14: Example of a Receiver Operating Characteristic Curve for an algorithm performing worse than random guessing

than 0.5, as in figure 3.14, then the model performs worse than random guessing and in fact reversing the outputs of the classifier, when classifier indicates 1 reverse it to 0 and when classifier indicates 0 reverse to 1, would yield better results in terms of the numerical value of AUC.

3.5 Cross-validation

The last theoretical aspect worth explaining in reference to this research thesis is the cross-validation. Cross-validation is a resampling method, which involves "repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model" (37).

One important reason for which cross-validation will be used in the evaluation procedure is to understand what is the variability of model's performance. The applicability of cross-validation is best explained by using an example:

1. Let's imagine that once the data has been cleaned and features have been extracted, the dataset is then divided into two parts as required for supervised machine learning: training set and test set.
2. Using the training set the model has been fit to the data and its performance has been measured on the test set. Subsequently this performance has been reported as the final performance of a given machine learning model. Such a method of evaluating model's performance is known as *validation set approach*.

The issue with validation set approach is that the model is trained only on a segment of entire dataset. This means that the model did not have the opportunity to capture all the trends in the data. Furthermore, under validation set approach only one score of evaluation metric is known. However, little information on the margins of those evaluation metric values are known.

In order to gain insight into the margins of the evaluation metric, cross-validation is used. The procedure for performing the cross-validation is the following:

1. The dataset is divided into 10 parts.
2. Subsequently, at every of the 10 iterations, nine of those parts were used as training set to fit the model and one of those parts was used as test set to evaluate the performance of the model.
3. Once all the 10 iterations have occurred, the mean and standard deviation/range of model's performance has been reported as final performance of a given machine learning model.

One can immediately notice the following benefit to using cross-validation for evaluating model performance: rather than reporting a singular value of a performance metric

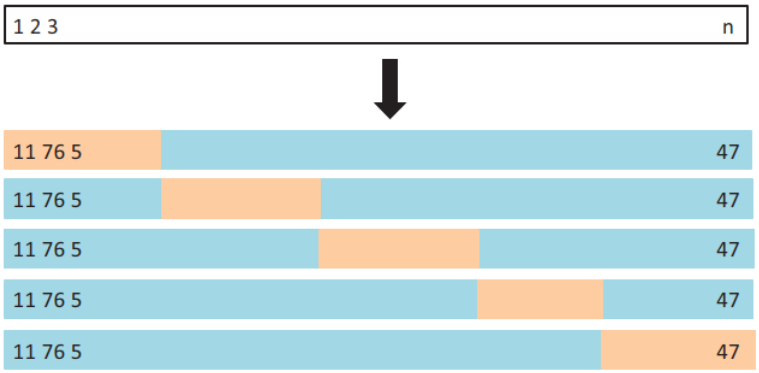

FIGURE 3.15: A schematic of a 5-fold cross validation. The white bar represents all observations in the dataset while the blue and orange bars represent the training and test sets respectively at every iteration. Source: (37)

of interest, which is a result of a singular random splitting of the dataset into test and training set, by using cross validation the performance's distributions is known, allowing thus to predict with greater confidence what the final evaluation metric value. Furthermore, by using cross-validation there is certainty that every observation was used in the test set, as opposed to the case without cross-validation where once the dataset is divided, only selected few observations are used to evaluate the performance metric on.

In reference to this master thesis, cross-validation is used on features derived from the Alliander dataset in order to select the optimal mix of hyperparameters for LR, DT and SVM. Subsequently validation set apporach is used for the three selected performing models where the training set is composed of the entirety of features derived from Alliander's load profiles and the test set is composed from the features extracted from simulated load profiles. The reason for selecting validation set approach is that by limiting the training dataset to Alliander's set of features only, insight into the potential for generalisation of models trained on in-house dataset is known.

3.6 Chapter summary

In this chapter LR was elaborated upon with details about logistic function and the maximum likelihood perspective given. Furthermore DT was discussed alongside the most popular splitting methods and the pruning procedure. Furthermore the notion behind SVM was elaborated by deriving it from Maximum Margin Classifiers and Support Vector Classifiers respectively. Lastly, evaluation criteria and cross-validation were discussed.

Chapter 4

Data

4.1 Chapter Introduction

This chapter will deal with the datasets there were used in the subsequent parts of the research. These datasets include load profiles of owners and non-owners of heat pumps provided by Alliander, heat-pump only profiles and load profiles of households without load profiles used for simulation of load profiles of heat pump owners as well as datasets containing temperature time-series data.

4.2 Load Profiles of Alliander

Alliander does not have access to load profiles of all households it delivers electricity to. In fact out of several million customers that this Distribution Network Operator supplied electricity to, only around 120k profiles are available.

Furthermore, finding households with heat pump is a difficult task. This stems from the fact that Alliander, the provider of the data, does not have access to information regarding the sales of heat pumps and cannot directly infer which household uses a heat pump. Thus in order to access load profiles of households with heat pump, which will be subsequently used to extract features for the labelled dataset, the following procedures were undertaken:

1. Areas within the country which are not connected to the gas network were identified - these stem from presence of modern apartment complexes, built in recent years, which do not use natural gas for heating.

2. These areas were cross-referenced with areas in which the 120k load profiles available at Alliander's database can be found.

3. Out of the cross-referenced profiles, 250 were randomly chosen.

In a similar fashion, the load profiles without heat pump were found, by excluding the areas of the country in which no gas network can be found. As a result of conducting the procedures outlined above, **both households with and without heat pumps were roughly evenly spread across Netherlands**. This has important considerations for the extraction of temperature data as discussed in the next section.

Load profiles extracted above have the following characteristics:

- Contain one year of data.
- Measured every 15 minutes in units of Watt-hours - thus there are 4*24*365=35040 observations per profile.
- Have an indication of whether the data was tempered with - more on it in the subsequent subsection.
- Have separate columns for net production and net consumption data, more on it in the subsequent section 4.2.1.

4.2.1 Smart metering

It is necessary to understand how do the smart meters deployed by Alliander work, in particular how are the measurements displayed in the data performed. This knowledge is necessary in order to grasp the information content of the data delivered by Alliander. A smart metering process can be seen on figure 4.1. On that image three plots can be observed:

- Plot 1 displays the instantaneous power of electricity consuming appliances at the household level, in other words the instantaneous power electricity consumption in red, and in blue the instantaneous power of electricity production appliances, assumed to be photovoltaic appliances. The polarity of instantaneous electricity production and consumption is opposite since one consumes electricity while the other produces it;
- Plot 2 displays the net electricity flow of the household calculated by adding the produced and consumed electricity. Since for the first 8 minutes greater amount

of electricity was produced than consumed, that net electricity flow is negative indicating the household delivers electricity to the grid. Subsequently, once the electricity consumed surpasses the produced one, the net electricity flow becomes positive, indicating the household is drawing electricity from the grid.

- Plot 3 displays the net electricity consumption and net electricity production. Net electricity production is positive when the net electricity flow is negative, as explained in the previous bullet point, while the net electricity consumption is positive when the net electricity flow is positive. It is important to note that within the 15-minute window, net electricity production and net electricity production are never below 0.

For every 15-minute interval, such as the one presented in figure 4.1, values of net electricity production and consumption are summed up and they are displayed in separate corresponding columns. Thus it is important to note that **in Alliander's dataset net electricity production and consumption do not represent absolute electricity consumption and production**. As a result it cannot be inferred from the data how much electricity was consumed or produced in a given day as the values at the disposal are net values.

More formally, the smart meter works in the following way:

1. for every minute within 15 minute interval - refer to plot 1 on 4.1:
   (a) for electricity production:
      i. integrate instantaneous power over the past minute to get total energy
      ii. record it.
   (b) for electricity consumption:
      i. integrate instantaneous power over the past minute to get total energy.
      ii. record it.
   (c) subtract total energy produced over the past minute from total energy consumed over the past minute - refer to plot 2 on 4.1, to get net electricity flow.
   (d) create two variables - net electricity production and net electricity consumption in the following way, refer to plot 3 on 4.1:
      i. if net electricity flow is greater than 0, assign the net electricity flow value to net electricity consumption. Assign 0 to net electricity production.
      ii. if net electricity flow is less than 0, assign absolute value of the net electricity flow to net electricity production. Assign 0 to net electricity consumption.

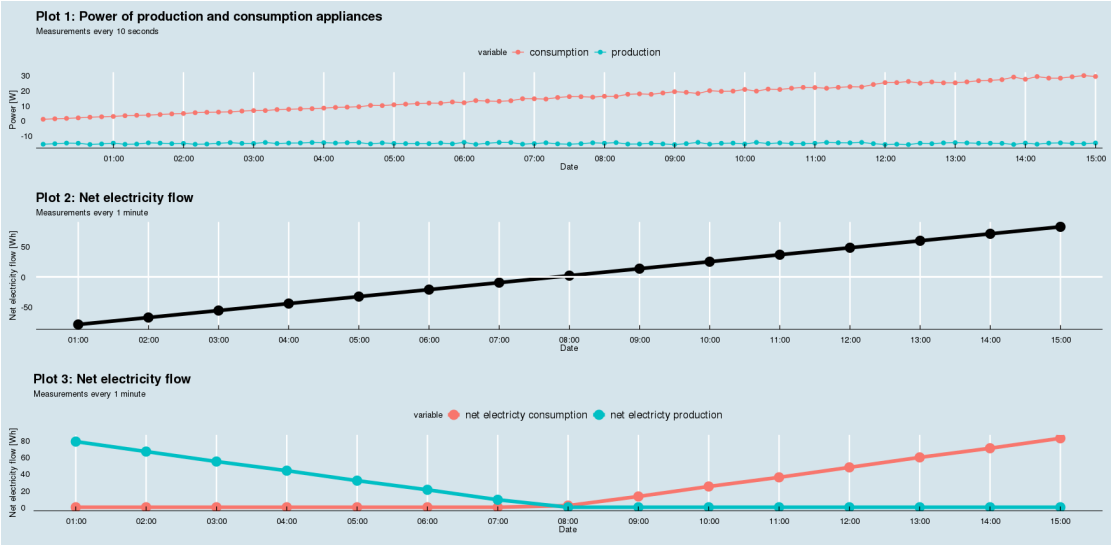

FIGURE 4.1: Power and electricity flow versus time

2. At the end of the 15 minute interval sum up net electricity consumption over the past 15 minutes. Output it to the final data.

3. At the end of the 15 minute interval sum up net electricity production over the past 15 minutes. Output it to the final data.

4.2.2 Overview of the data

Since it was assumed, that anyone wanting to replicate the method for heat pump detection presented in that thesis has only net electricity flow data, the net electricity consumption was subtracted from net electricity production yielding net electricity flow, as presented on plot 2 of 4.1. **The features presented in subsequent sections will be calculated based on net electricity flow values.**

Broadly speaking, not only owners and non-owners of heat pumps can be distinguished in the dataset, but also owners and non-owners of photovoltaic appliances, which produce surplus energy in summer days. Examples of every such user can be found in figures 4.2 and 4.3.

Looking at the overall distribution of load profiles among those with and without heat pump we can observe that the owners of heat pumps tend to have a higher median consumption throughout the entire year, as seen on figures 4.4 and 4.5.

The dataset also contained load profiles with missing data, as presented on figure 4.6. It can be seen that right before we can observe a string of missing observations, around mid April, there is a spike in the energy consumption. In order to understand why

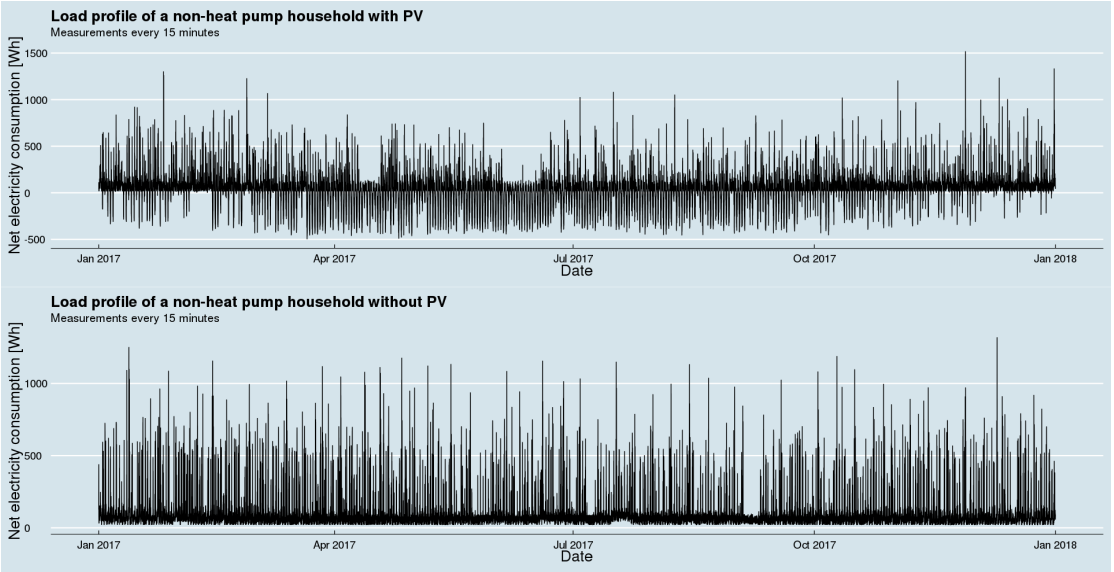

FIGURE 4.2: Net electricity flow of households without heat pump for households with and without photovoltaic appliances

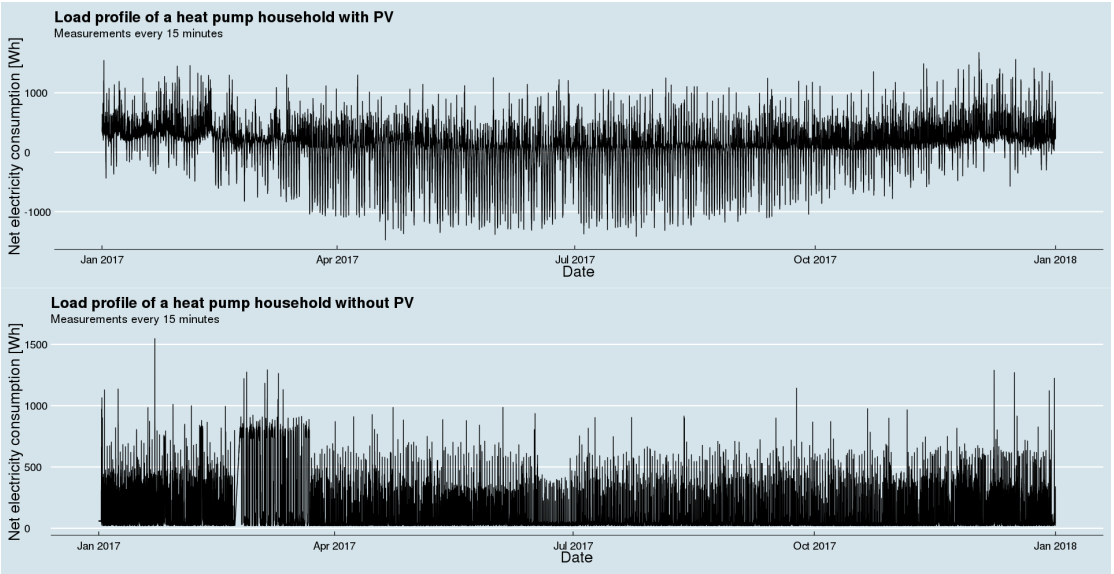

FIGURE 4.3: Net electricity flow of households with heat pump for households with and without photovoltaic appliances

does this spike occur, it is necessary to delineate the procedures activated by the smart meter communication system when a missing data is detected. Using an example:

1. If for 30 minutes, say from 01:30 through 01:45 and 02:00 the meter does not communicate with the database, the above-mentioned observations get missing values observations missing values for both net electricity consumption and production, and thus for net energy flow too, as depicted on plot 1 of figure 4.6.

2. Then at 02:15 the communication resumes and a sudden spike occurs, assigned by the system to the earliest non-missing available point, that is 01:15 before the

FIGURE 4.4: 90th and 10th percentiles as well as median of electricity flow of load profiles of heat pump non-owners

FIGURE 4.5: 90th and 10th percentiles as well as median of electricity flow of load profiles of heat pump owners

outage occurs. This spike represents the sum of all the points with the missing data, as shown on plot 2 of figure 4.6.

3. Observation with the spike is assigned as non-missing while the other are designated as missing values.

4. (Optional step) subsequently the data gets interpolated in a way that it averages for all the missing observations the value of energy consumption as in plot 3 of figure 4.6

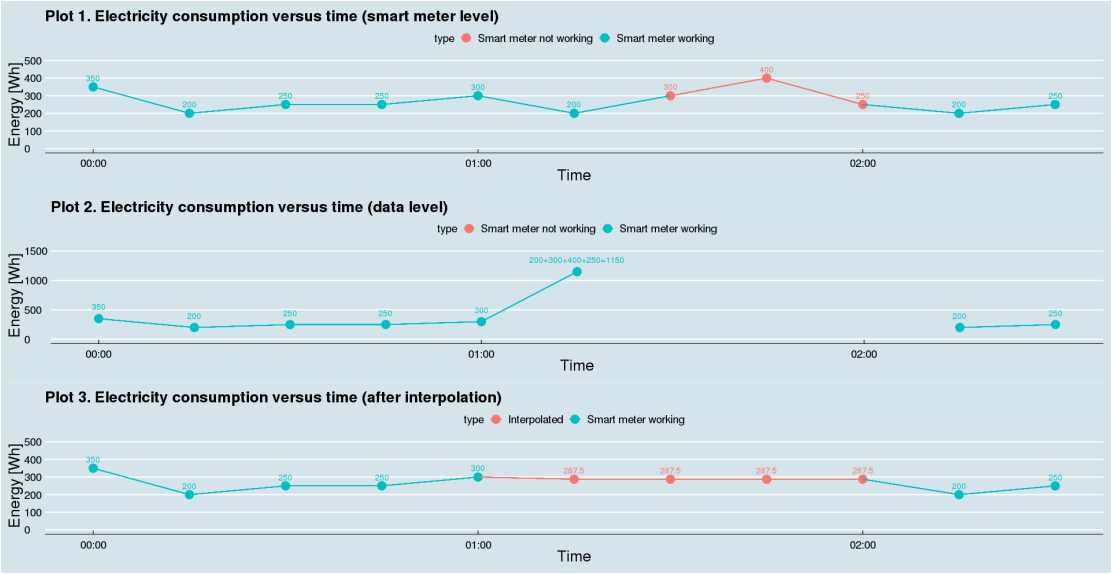

FIGURE 4.6: A load profile with missing data

In order to deal with those observations it was first inspected what share of all values do the missing values represent. They represent some 2.5% of all values. Next, it was calculated whether any of the profiles has more than 5% of missing values - and no such profiles exist. Lastly, since the missing data represented a relatively small share of all values for all profiles, these observations were interpolated as explained in the fourth point above.

4.3 Heat pump only load profiles

As mentioned in the introduction chapter, load profiles of owners of heat pumps will be simulated. This will be accomplished by pairing up load profiles of heat-pumps only with load profiles of households without heat pump - more on simulation technique in chapter 5.

The data for heat pumps was sourced from (41). The dataset contained 418 heat pump-only load profiles, that is time series electricity consumption data. The data was collected every 2 minutes. The period of measurement varied between 10 months and one year. Furthermore, the time period of the measurement was not uniform across all heat pumps with some of them recording measurements as early as 2012, while the other ones starting only in 2014. Furthermore, just as it is the case with the Alliander dataset, the exact location of each heat pump is not known and thus it is impossible to extract exact outside temperature - more on that in section 4.5.3.

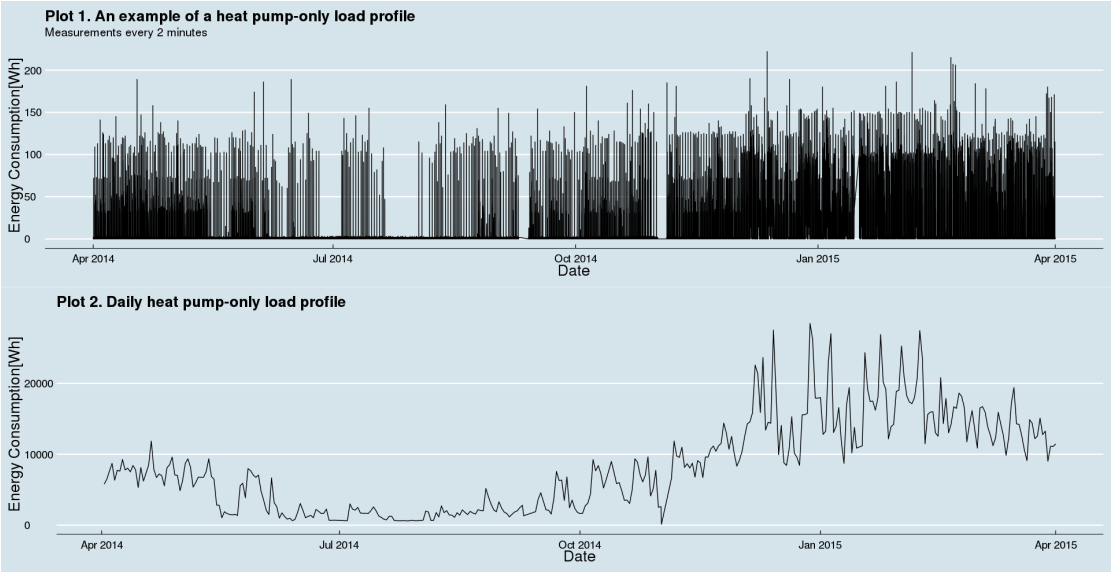

FIGURE 4.7: A load profile of heat pump only

The dataset is not of high quality. This is why certain criteria were established in order to determine whether a given heat pump load profile will be used in simulation. These criteria were the following:

- at least 363 days of data were required, seeing that this amounts to almost one full year and most of the data in the dataset had indeed 363 days of measurements;
  - Side note: 363 days may seem to be a arbitrary choice. If one full year was required 365 days should have been chosen. However, only 2 profiles actually had 365 days of data. Most had 363 days of data available thus 363 days was chosen as lower limit. Of course this distorts values of some of the features, but it does so by a very small margin - by a margin of 3 days only. This loss was deemed acceptable for the sake of obtaining enough heat pump load profiles.
- a heat pump load profile cannot have less than 90% of non-missing data, to ensure reliability of the energy consumption metrics;
- there could not have been more than 7 days of continuously missing data, as interpolation of missing data across more than a week made the load profile unreliable;

After the filtering, 320 load profiles of heat pumps-only have remained. Median and percentile plots of heat pump profile time series can be seen on 4.8

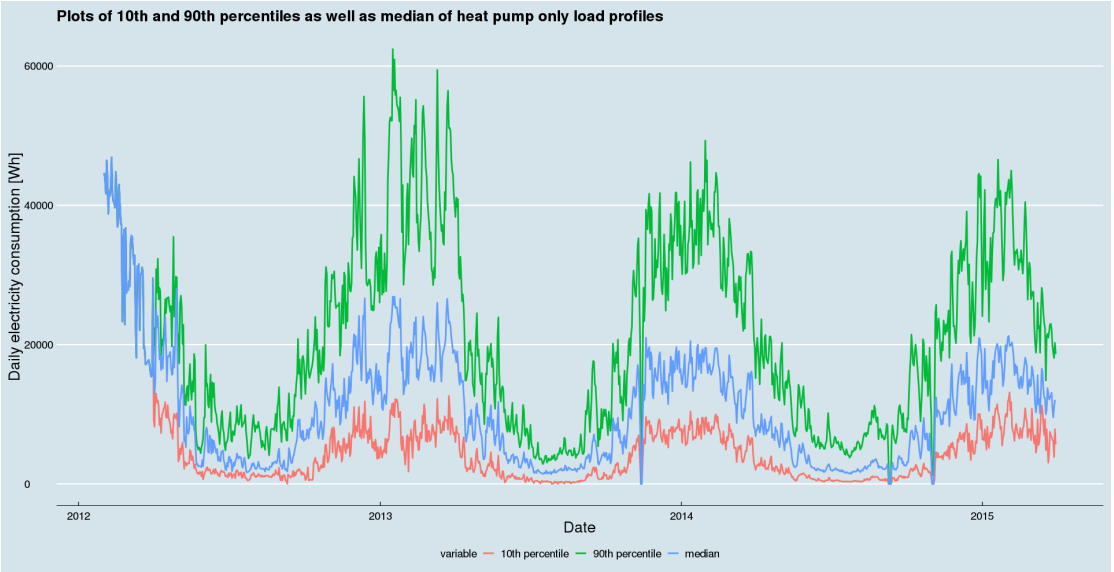

FIGURE 4.8: Plot of 10th and 90th percentiles as well as median electricity consumption of heat pumps

4.4 Load Profiles for simulation

The other crucial component needed to simulate load profiles of households with heat pumps are the load profiles of households without heat pumps. At the same time those profiles, which will be called *baseload* profiles in the subsequent chapters, cannot be from any database which might belong to Alliander, since the whole point of simulation is to provide and independent validation set. Such a database of baseload profiles was found at (42), provided by the UK Power Networks, one of the DNOs of the United Kingdom. An example of baseload profile can be see at 4.9

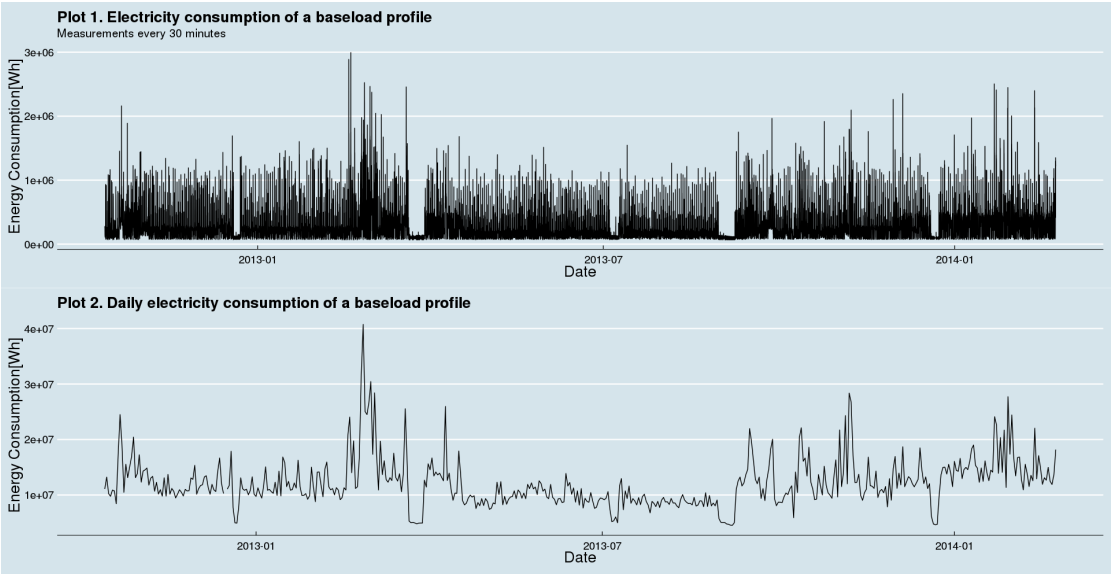

FIGURE 4.9: A load profile of a household from London Low Carbon dataset

In that dataset baseload profiles of over 5567 households can be found. It is important to mention however, that not all baseload profiles could have been immediately taken for the simulation. The reasons for that are the following:

- One-fifth of all households that have participated in an experiment whereby they were offered a demand-response tariff, that is one which changes depending on the time of the day. These users were filtered out, since as of the moment of this writing there are no records of Dutch customers participating in the same tariff schemes. Furthermore, changed tariffs had a purpose of lowering electricity consumption at certain times of the day, thus taking those households into consideration in this research might add two completely different groups in the baseload profiles set.
- As in case of heat pumps data was filtered out to ensure at least 90% of observations are non-missing, there is never more than one week of continuously missing data and there had to be at least one year of records;
- Furthermore, only those profiles which have records from January until December of 2013 of any given year were accepted (This criteria was not included in the case of heat pump only load profiles, due to scarcity of those as compared to over 4k available baseload profiles);

Having applied all of the above-mentioned criteria to baseload profile selection out of initial 5567 load profiles, 880 remained. The plot of median, 10th and 90th percentiles of electricity consumption of baseload profiles is displayed on figure 4.10

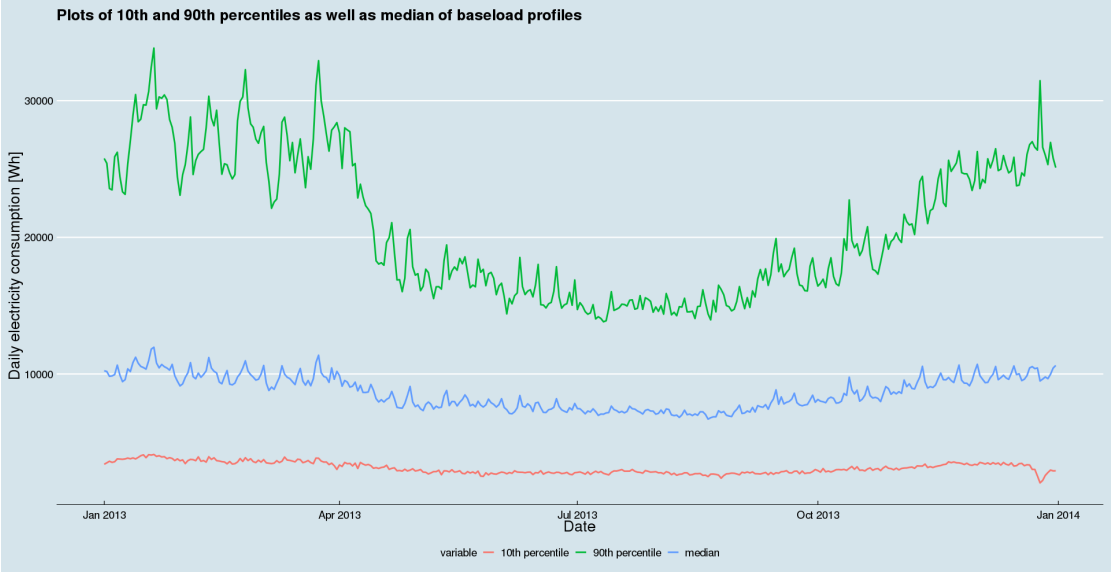

FIGURE 4.10: Plot of 10th and 90th percentiles as well as median electricity consumption of baseload profiles

Thanks to detailed information regarding the appliances used by each customer in the baseload profiles dataset, it was concluded that none of them used heat pumps. However, it can be seen on the plot above, the general trend is that electricity consumption increases in colder months. This might indicate that some of the baseload profile users use electric heating as heating source.

4.5 Temperature data

In order to extract features associated with temperature, more on that in section 6.2, temperature records had to be obtained. There are two main aspects of extracting necessary temperature data: the right period and the right location. Temperature data depending on their location will be discussed below. Each of the described temperature sets will be paired with a load profile dataset and thus appropriate period of temperature records will be kept.

4.5.1 Dutch temperature data

The the temperature data, provided by the Royal Netherlands Meteorological Institute - *Koninklijk Nederlands Meteorologisch Instituut*, were found at (43). The load profiles provided by Alliander used as input for feature calculation in the first part come from 2017, thus temperature data from that year was used. The biggest issue however was the fact that the exact locations of households of which load profiles were at disposal were not known. Thus it was assumed that the load profiles in both datasets were spread equally geographically within Netherlands, and thus taking the temperature readings from De Bilt, which is city located in the centre of the Netherlands, will give a good approximation of the overall temperature.

4.5.2 London temperature data

As far as the load profiles from the London Low Carbon dataset, discussed in 4.4, are concerned, since the Location was known, London, and the year of profiles is also known, 2013, appropriate temperature data was extracted from rp.ru which provides free archive weather information for locations all over the globe. The plot of mean daily temperature in London can be seen at 4.12.

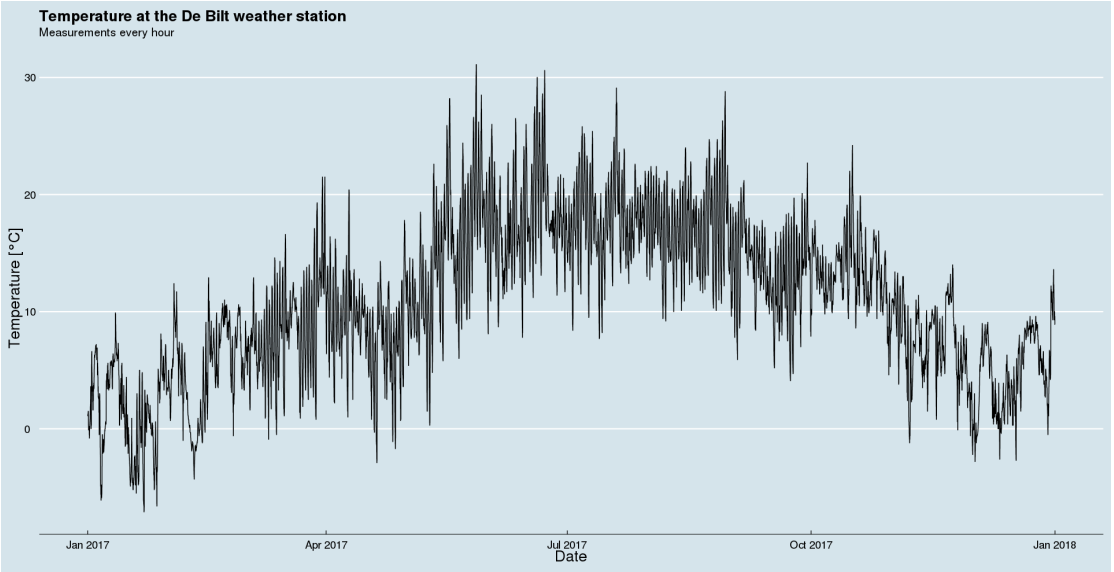

FIGURE 4.11: Temperature readings of De Bilet weather station. Source: (43)

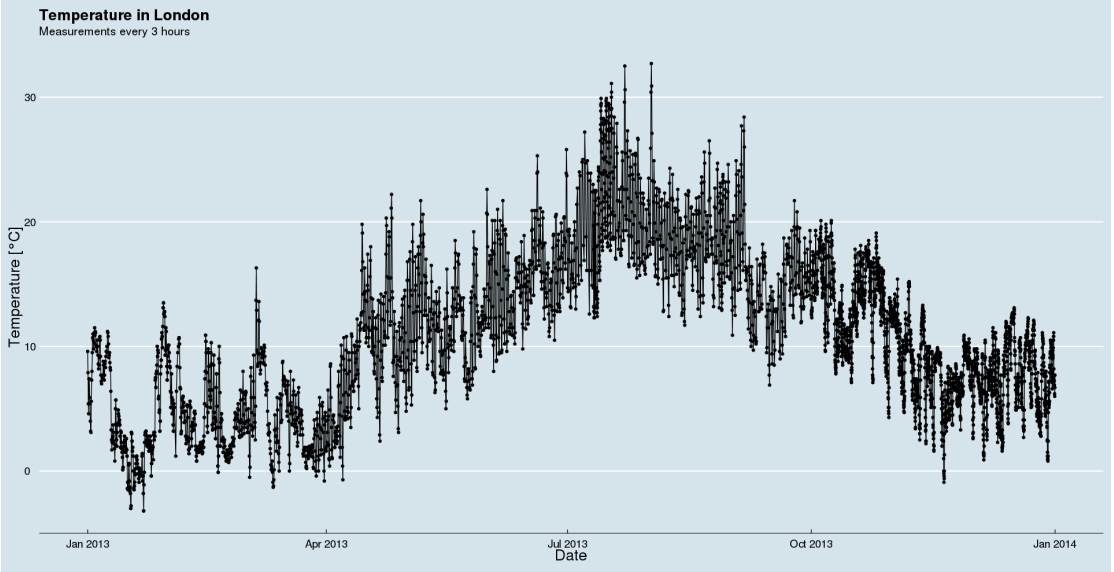

FIGURE 4.12: Temperature from London weather station. Source: rp.ru

4.5.3 Nottingham temperature data

In a similar fashion to the load profiles provided by Alliander, the heat pump load profiles do not originate from a single location. Furthermore, the exact list of locations is not known. The only given information is that the heat pump load profiles originate from the United Kingdom. Thus in order to extract features associated with temperature, temperature data was taken from weather station which was as close as possible to the United Kingdom's population centre, which is located in the village of Appleby Parva (44). The nearest available weather station is located in Nottingham and once again the temperature data of Nottingham available at rp.ru wes extracted. The time

period ranged from 1st of February 2012 till 31st of March 2015, which corresponds with time span of the heat pump load profiles.

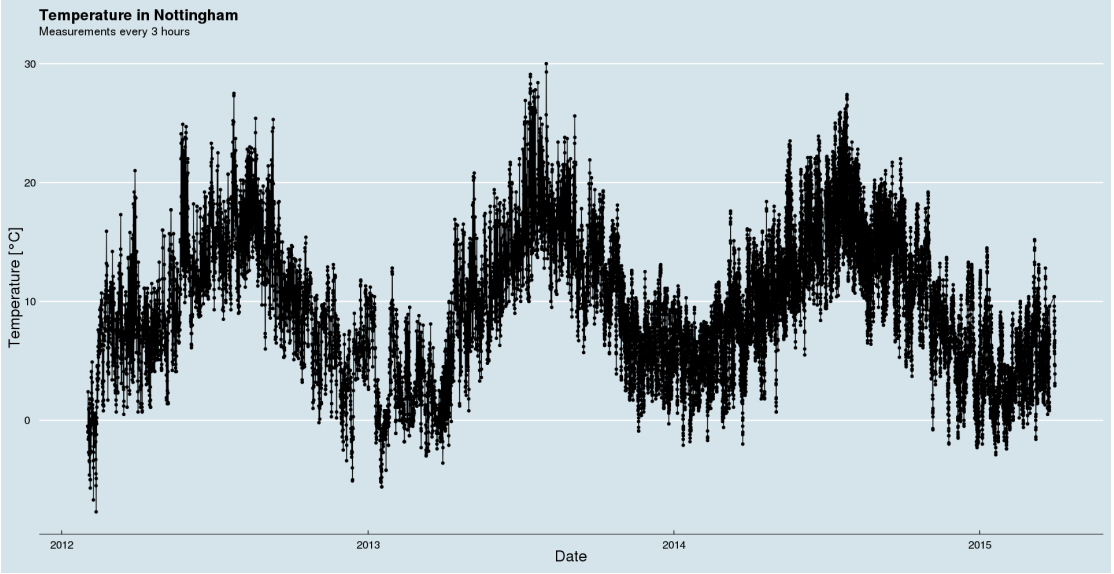

FIGURE 4.13: Temperature from Nottingham weather station. Source: rp.ru

4.6 Chapter summary

In this chapter the raw data used at different stages of the research project was discussed and visualised. First the smart metering system operating principles were discussed with detailed elaboration on how missing data was handled. Next, visualisations of exemplary load profiles with and without heat pump and PV appliances were demonstrated. In similar fashion, datasets used for simulating load profiles of households with heat pumps were discussed. Lastly, the issues regarding the temperature data were elaborated on.

Chapter 5

Simulation and analysis of new load profiles

5.1 Chapter Introduction

This chapter describes the simulation process of load profiles of households with heat pump. Next, analysis of those simulated load profiles is performed and comparisons are drawn between the simulated load profiles and load profiles of heat-pump owners supplied by Alliander

5.2 Simulation process

Creation of synthetic load profiles of households with heat pumps requires two components: load profiles of households without heat pump (baseload profiles) - such households heat their spaces with gas network - and load profiles of heat pump devices only of which there are 880 and 320 respectively. By the end of the simulation process there will be 880 simulated load profiles of heat pump owners.

There are two aspects associated with the simulation process. The first one is the aspect of handling differences in profiles lengths. The other corresponds to pairing of heat pump only load profiles to baselaod profiles.

5.2.1 Handling profiles differences

As mentioned in the previous chapter, heat pump only load profiles can have only 364 or 363 days of data as opposed to baseload profiles having 365 days of data. Furthermore, baseload profiles are all obtained from year 2013, while heat pump only load profiles span from 2012 to mid-2015.

The following method was devised to solve the problem of misalignment and difference in length among the profiles:

- If heat pump only load profile spanned across two years, say 1st June 2012 to 29th May 2013, corresponding baseload profile time indexes were mapped on time indexes of the heat pump only load profile in the following way:
  1. the year component of baseload profile was dropped
  2. the period between day and month of the first time index in heat pump only load profile (in the example above 1st June) and 31 December was extracted from the baseload profile and the year corresponding to the earliest year in heat pump only load profile was assigned to it - 2012 in the example above.
  3. the period between 1st January of baseload profile and the end day and month of heat pump only load profile (29th May in the example above) was extracted from the baseload profile and the later year of the heat pump only load profile was assigned to it - 2013 in the example above.
- If heat pump only load profile spanned across on year, the year of the baseload profile would be changed to correspond to the year of the time index of heat pump only load profile and an inner join on time indexes would occur.

The method above thus ensures that baseload profile, which does not exhibit strong dependence on temperature, is mapped on heat pump only load profile, which does exhibit strong dependence on temperature. This is why the time indexes of baseload profiles were adjusted to match time indexes of heat pump only load profiles and not the other way around.

5.2.2 Pairing the profiles

As far as pairing of the profiles is concerned, the following aspects need to be taken in consideration:

1. heat pump owners who posses larger households have heat pumps which consume larger quantity of electricity to heat up the increased living space. Similarly, if a heat pump is used to provide heating for a small apartment, it consumes considerably less energy;

2. there has to be a large enough quantity of both heat pump profiles and baseload profiles in order to create a dataset of diverse load profiles of households with heat pump owners in order to ensure that the final machine learning algorithm generalises well for the entire database of Alliander's load profiles;

As far as ensuring that there is a large enough of pool of heat-pump only load profiles as compared to number of baselaod profiles is concerned, it occurred not to be a problem since in total having filtered out the original raw data there were at disposal 881 baseload profiles and 320 heat pump only load profiles. This means that for every heat pump only load profile there were 3 baseload profiles, which is a sufficiently high number to ensure a diversity in final simulated load profiles of heat pump owners.

From the first concern mentioned above, it stems that baseload profiles which consume largest amount of electricity throughout year should be paired with heat pumps which consume largest amount of electricity per year. In more formal terms, the pairing of baselaod profiles and heat pump profiles should be based on rank of those profiles corresponding to the relative electricity consumption of all baseload and heat pump profiles respectively. In more formal terms the following procedures were applied for the pairing mechanism:

1. Rank in descending order all heat pump only load profiles in accordance to total yearly electricity consumption, with 1 being the highest yearly electricity consumption;

2. Rank in descending order all baselaod profiles in accordance to total yearly electricity consumption, with 1 being the highest yearly electricity consumption;

3. For i in ranks of heat pump only load profiles:

   (a) pair heat pump only load profile with rank==i with baseload profiles ranked i, i-1 and i-2.

An example of this pairing mechanism, with 3 heat pump only load profiles and 9 baseload profiles, can be seen on 5.1. Having paired the heat ump only load profiles and baselaod profiles using the above-mentioned pairing scheme, 881 simulated load profiles of households with heat pumps were created.

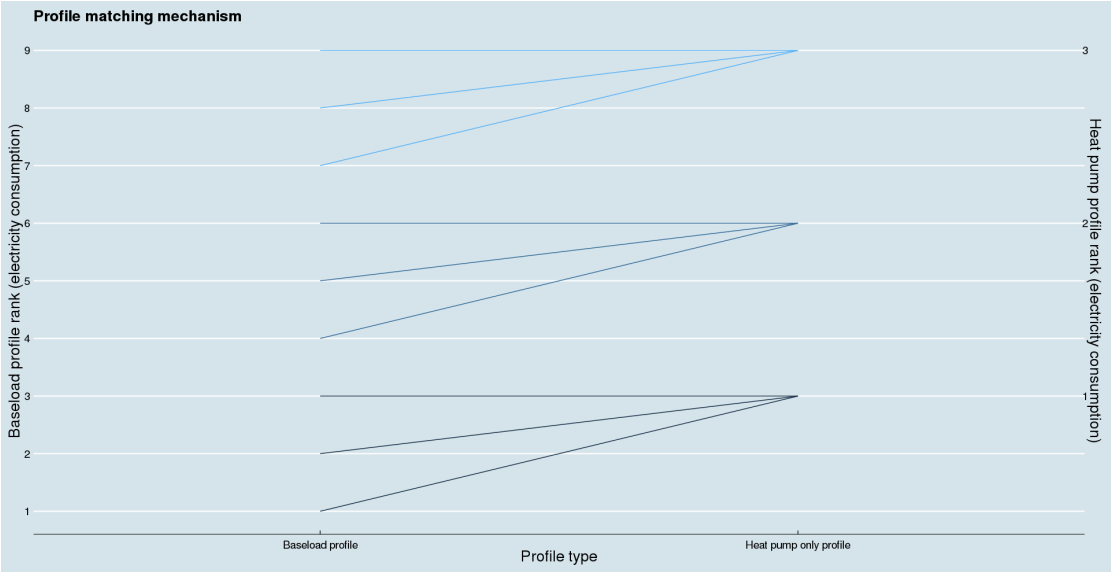

FIGURE 5.1: A visualisation of matching scheme used for pairing heat pump only load profiles with baselaod profiles

5.3 Simulation output analysis

Once the profiles have been simulated, they need to be analysed. An example of a simulated load profile is seen on 5.2. As it can be seen by comparing figure 5.3 which shows the median and percentile plot of simulated load profiles and figure 4.8 which shows percentile and median plots of heat pumps owners, the simulated data, as expected, does not contain any PV production, since it was simulated by using load profiles with electricity consumption only.

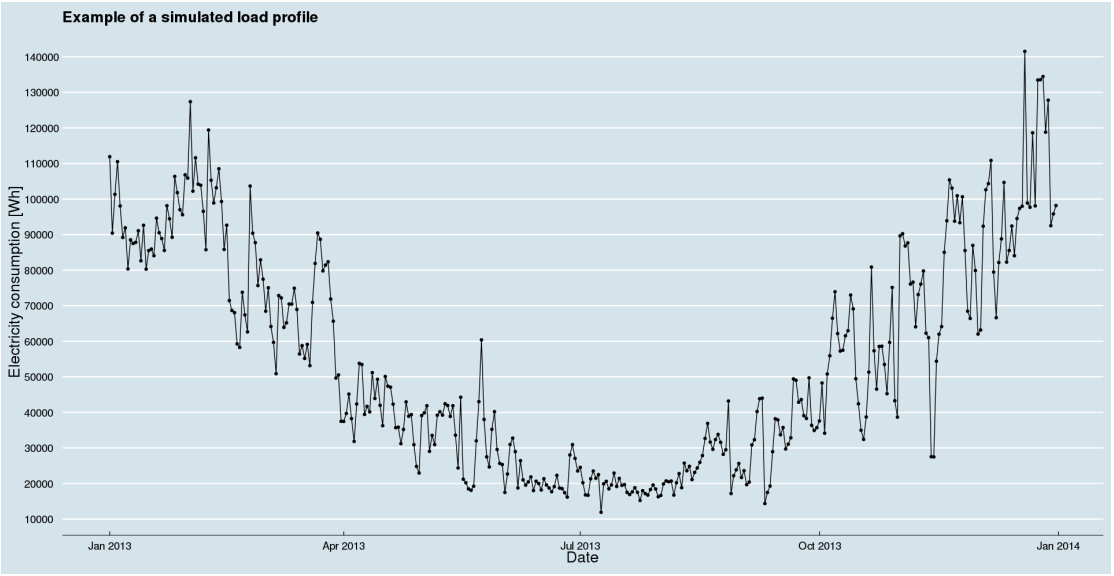

FIGURE 5.2: An example of a simulated load profile

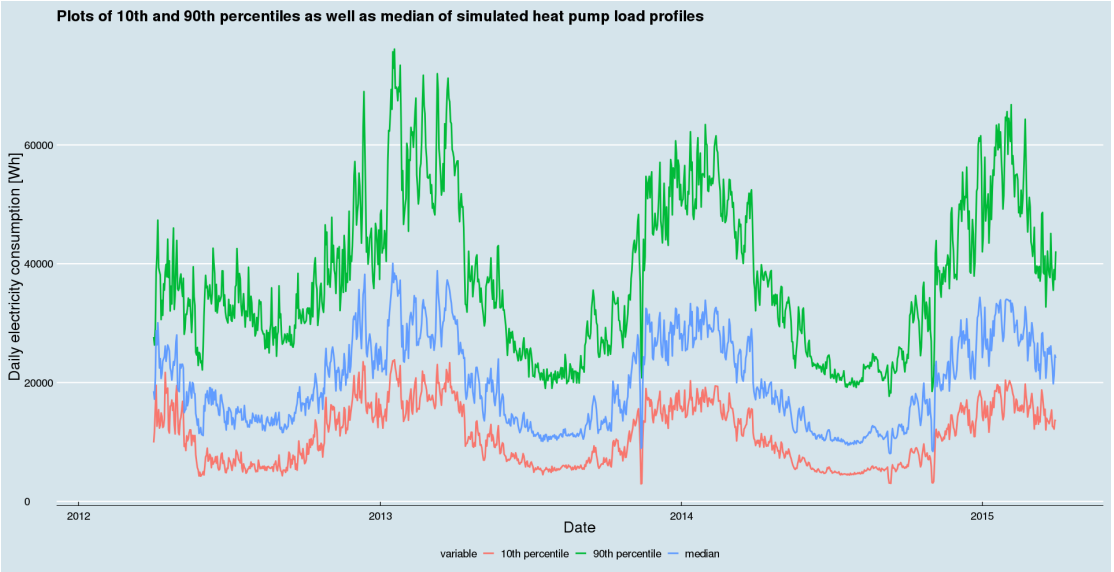

FIGURE 5.3: Median, 10th and 90th percentile plot of simulated data

All in all the percentile and median plot exhibits expected behaviour since the consumption increases in winter months and decreases in summer months. The fact that the consumption decreases in summer months does prove an important point: heat pumps in the dataset of heat pump only load profiles are not used during summer as air conditioners, which stands contrary to what has been reported so far in literature (30).

5.4 Chapter summary

In this chapter the mechanism for simulating load profiles of heat pumps owners has been demonstrated. Subsequently analysis has been performed on the simulated data and conclusion was drawn that the dataset exhibits expected properties, thus showing that the simulation process was suitable.

Chapter 6

Features

6.1 Chapter Introduction

In this chapter the reasoning behind choosing specified set of features is provided. Furthermore, the numerical values of those features calculated for both, the Alliander-based dataset and the simulated one are analysed.

6.2 Feature choice

Time series data contain numerous observations by definition. For that reason, it would be unwise to use all of them as separate features, since this might lead to overfitting. This is why features, that is characteristics describing load profiles, need to be extracted and subsequently used as input to machine learning. Since this master thesis goals revolve around using machine learning models to distinguish among owners and not of heat pumps, understanding which characteristics pertaining to electricity consumption differ among the two groups, is crucial. The goal is to capture those characteristics in form of numerical or categorical features. Research which was focused on similar problem of heat pump detection using smart meter data (30) used the following features:

- temperature dependent heating parameter - the slope between mean daily temperature and daily electricity consumption during winter months;
- temperature dependent cooling parameter - the slope between mean daily temperature and daily electricity consumption during summer months;
- ratio between the average energy consumption in winter and average energy consumption in spring and autumn;

- average energy consumption during winter;
- average energy consumption during spring and autumn;

The reasoning behind choosing this particular set of features is the following:

- for features 1 and 2: it is expected the heat pump is used as both air conditioner and source of heating. Thus during summer and winter the energy usage of the entire household should be correlated with the temperature in the following way:
  - The higher the temperature in summer the higher electricity consumption since the need for air conditioning increases;
  - The lower the temperature in winter the higher electricity consumption since the need for heating increases;
- for feature 3: it is expected that during spring and autumn there is no need for either heating or air conditioning thus if the average energy consumption in winter - which ought to be high for households with heat pump - was divided by the average energy consumption in spring and autumn, clear differentiation between two groups would be made in the following manner:
  - for heat pump owners this ratio is high since there is a large difference between electricity consumption in winter and spring/autumn due to the fact that heat pumps are used in winter as opposed to spring and autumn;
  - for households without heat pump this ratio is close to one due to the fact the heating is not powered by electricity - this group would thus represent households without heat pump.
- for feature 4: the average energy consumption during winter would be high for owners of heat pumps since heat pumps are used for heating while it would be low for households without since heating is powered by gas, which does not draw electricity;
- for feature 5: this feature would not differentiate well between the owners and non owners of heat pump since spring and autumn is the period of the year when heat pumps are not used;

The above-mentioned features cannot be however used in this research for the following reasons:

- In the dataset provided by Alliander several confounding variables are present such as the size of houses: most of the households with heat pump in our dataset

are stand-alone homes while the ones without are apartments. Using average energy consumption thus would not only indicate larger electricity consumption due to heat pump presence but also due to large house size;

- In (30) the authors had at their disposal electricity consumption data, rather than electricity flow data, as is the case in this research. Repeating identical calculation methods would cause distortion since electricity production would be included, as it can be seen in figure 4.5;
- For feature 1 and 2: calculating the slope is not the optimal way of differentiating owners and non owners of heat pumps since slope does not describe the impact of temperature's change on daily energy consumption. Other metrics such as coefficient of determination better capture the impact of temperature change on energy consumption.

Solving the first problem, which is the presence of PV production in Alliander's load profiles, requires extraction of energy flow data of periods during the day in which no PV production is present. An obvious choice is to extract load profile observations which occurred only during the night - defined as any observations that occurred between 10 pm and 4 am when there is no sunlight. It can be noticed that in vast majority of load profiles in Alliander's dataset, which contains electricity flow data, no production is present by inspecting the figure 6.4. Of course this problem is nonexistent in case of simulated datasets since those were extracted from consumption-only data. Nevertheless, for continuity, those datasets as well will be limited to night-time observations only.

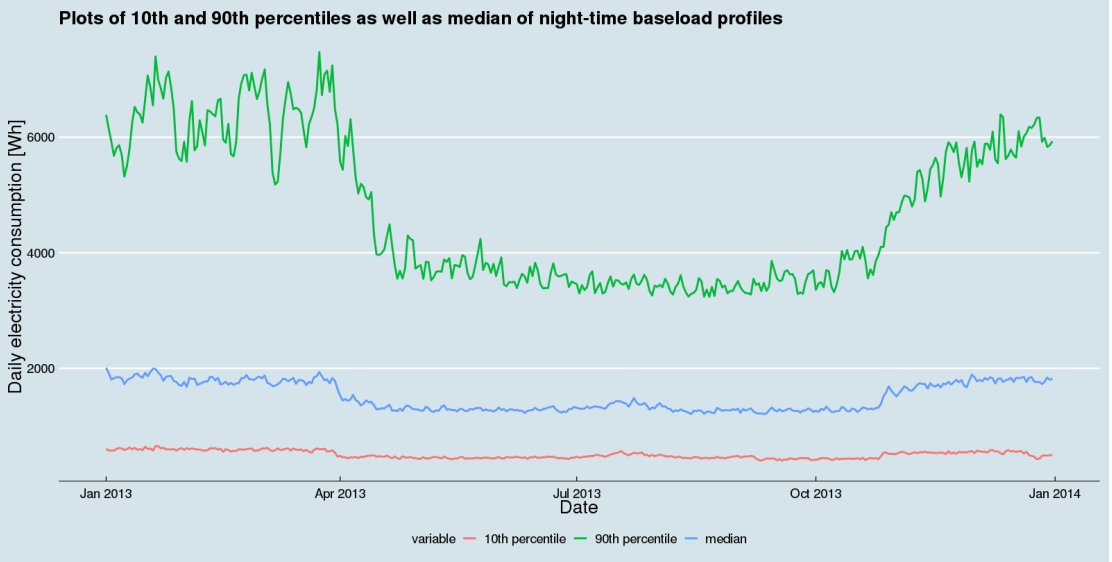

FIGURE 6.1: 90th and 10th percentiles as well as median of electricity consumption of baseload profiles

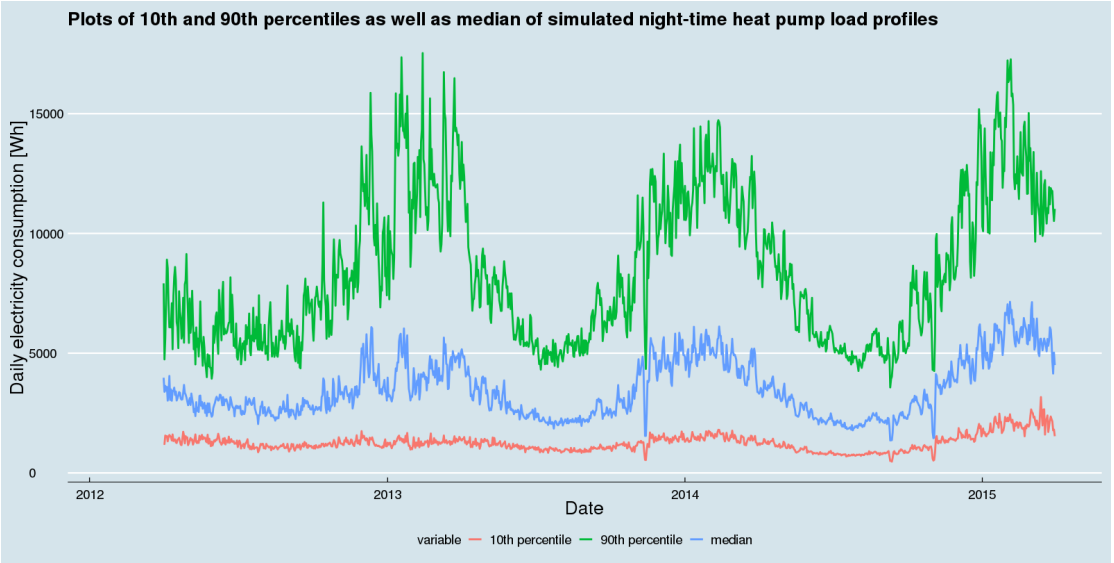

FIGURE 6.2: 90th and 10th percentiles as well as median of electricity consumption of load profiles of simulated heat pump owners

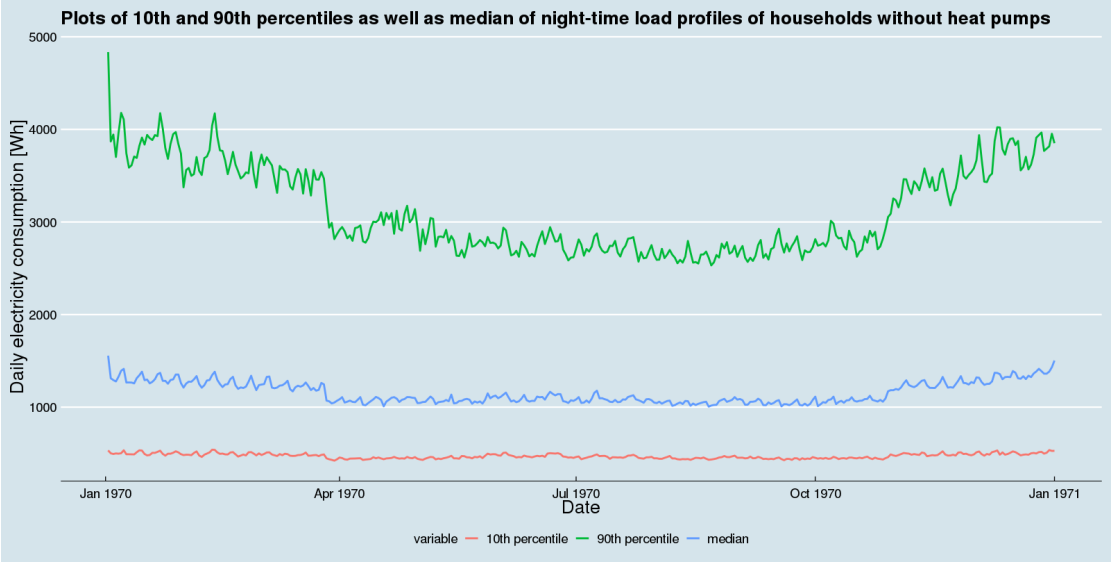

FIGURE 6.3: 90th and 10th percentiles as well as median of electricity consumption of load profiles of heat pump non-owners from Alliander's set

Furthermore, in order to improve the feature set, another statistical parameter which explains the proportion of variance in daily electricity consumption explained by variance in between mean daily temperature, namely the coefficient of determination, will be extracted from each profile. It takes values ranging from 0 to 1 and it quantifies proportion of the variance in the dependent variable, daily electricity consumption in this case, that is predictable from the independent variable, mean daily temperature. The mathematical expression for coefficient of determination is the following:

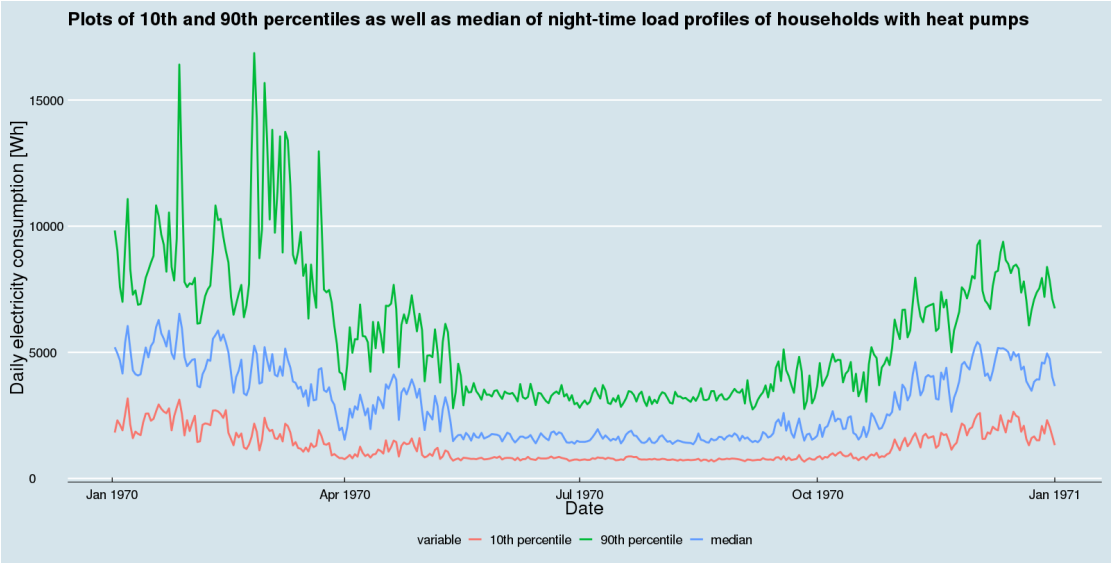

FIGURE 6.4: 90th and 10th percentiles as well as median of electricity consumption of load profiles of heat pump owners from Alliander's set

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where n is the total number of observations in the set, y_i is the actual value of the label at instance i, f_i is the output of the machine learning models at instance i, and $\bar{y}$ is the average value of the label.

To demonstrate the use case of coefficient of determination one should look at figure 6.5. In that figure the plots of mean daily temperature, x-axis, and daily electricity consumption, y-axis, for owners and not of heat pumps from the Alliander dataset are presented. It can be seen that in both cases if a least-square regression line was fit, its slope would be negative. But another feature which would allow for even greater degree of differentiation is coefficient of determination, which rather than explaining whether the relation between mean daily temperature and electricity consumption is positive or negative, it allows to understand what percentage of variance of electricity consumption is explained by variance in mean daily temperature. As it can be seen on figure 6.5 the slopes for owners and not of heat pumps are equal to -0.027 and -0.019 respectively, difference of 43% while coefficients of determination are equal to 0.619 and 0.326, difference of 190%. Thus coefficient of determination will be used alongside the slope values as a feature, since they allow for another degree of seperation between the two groups.

In order to ensure that our features would be empirically sound and would differentiate well between owners and not of heat pumps without accounting for confounding variables such as household size, all the features were calculated from range normalised

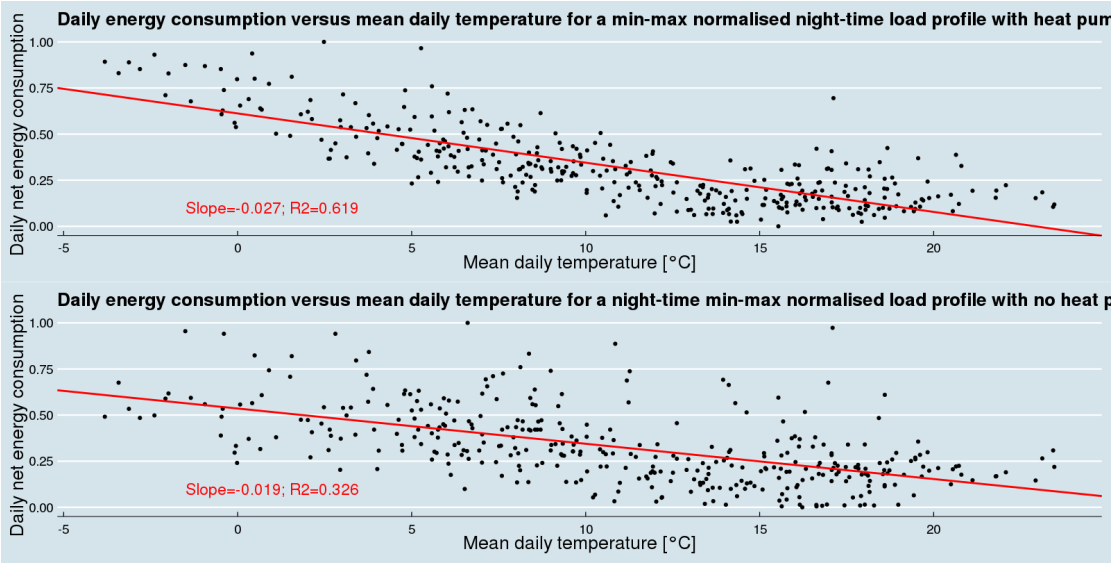

FIGURE 6.5: Net daily night-time energy consumption versus mean daily temperature for random households with and without a heat pump after min-max normalisation

load profiles. Range normalised load profiles are manipulated by subtracting from each observation within a load profile the minimum value of each load profile and dividing it by the range of values within that profile. This ensured that each load profile is within the 0, 1 range. The formula for an individual load profile min-max normalisation is the following: $X_i' = (X_i - X_{min})/(X_{max} - X_{min})$
where i represents an observation within a profile X. This ensures that the scale of all load profiles is identical thus 'bypassing' the confounding variable of higher absolute energy consumption of households with heat pump, as shown on figure 6.6

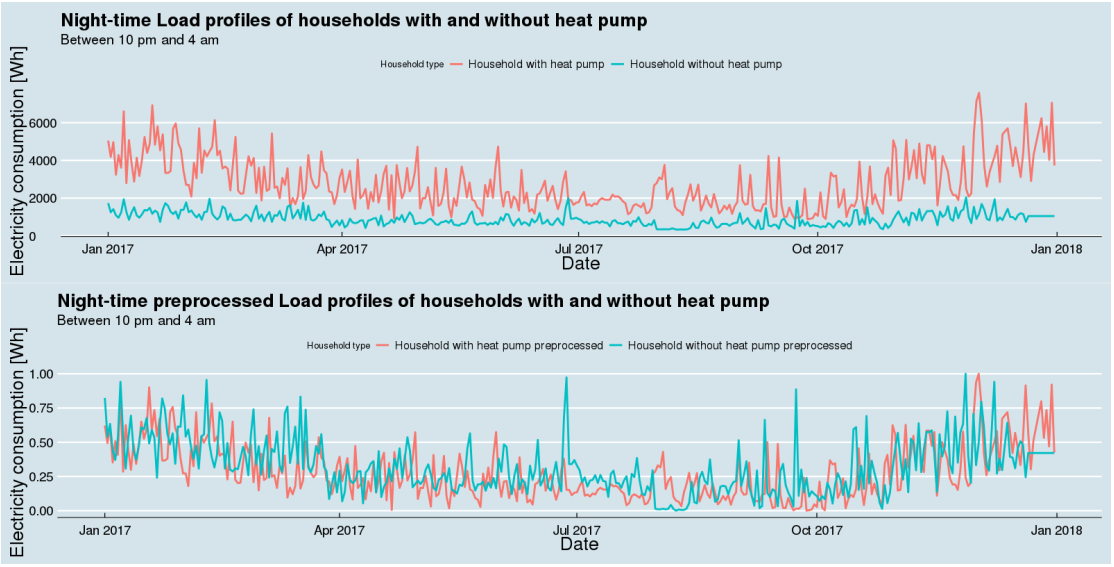

FIGURE 6.6: Examples of night-time original and normalised load profiles of households with and without heat pump

The following features were extracted and will be used as input to the machine learning algorithms:

1. **average night-time energy consumption** of a min-max normalised profile during the months of December and January;
2. **ratio** between average night-time energy consumption of a min-max normalised profile during the months of December and January and average night-time energy consumption of a min-max normalised profile during months of July and August;
3. **temperature dependent heating parameter** during the night - slope of the curve where x-axis is the mean daily temperature during night hours while y-axis is the daily night-time energy consumption of a min-max normalised profile;
4. **coefficient of determination** of the relation between mean temperature during the night and average consumption during the night - that is the coefficient of determination between the curve where x-axis is the mean daily temperature during night hours while y-axis is the daily night-time energy consumption of a min-max normalised profile;

6.3 Feature analysis

Conditional box plots of features for Alliander's and simulated dataset can be seen on figures 6.7 and 6.8 respectively. Next, each feature set will be analysed.

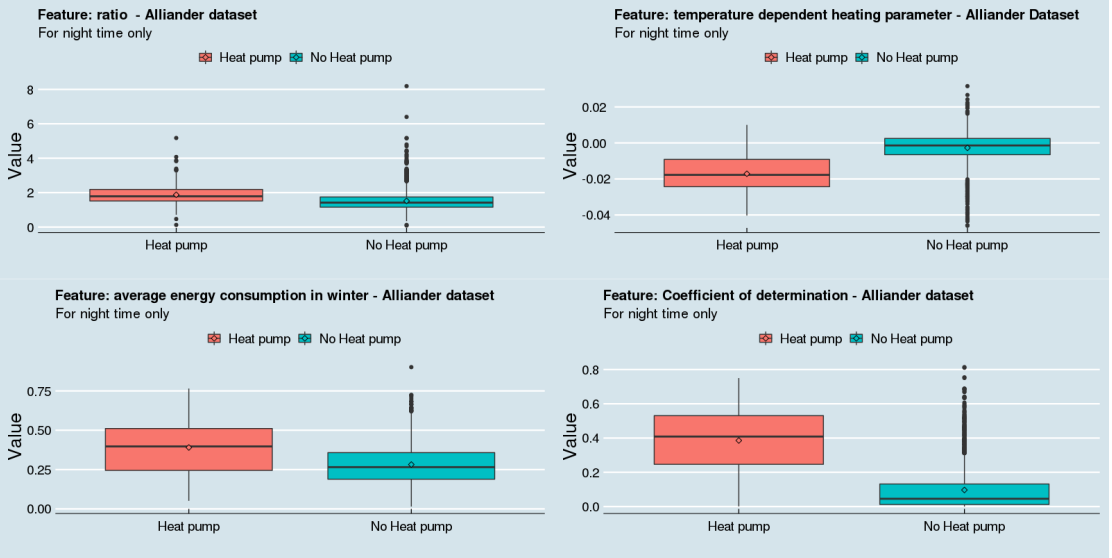

FIGURE 6.7: Conditional box plots of the features for Alliander dataset

As it can be seen by analysing the Alliander-based features, these seem to provide good degree of separation between the owners and non-owners of heat pumps. The coefficient of determination has significantly higher median value for owners, which is as expected considering that temperature does impact electricity consumption for heat pump owners.

Another feature with clear separation is slope. The median value for the slope for non-owners is very close to 0, signifying there is no relation between electricity consumption and changes in mean daily temperature. This stands at stark contrast to the median values of slope for heat pump owners. It is worth mentioning that while the absolute values of slope are small, reaching maximum 0.04, this is because the slope was calculated for min-max normalised profiles where the range for each load profiles is between 0 and 1. As an example for a slope value of -0.02 and the range in mean daily temperatures throughout the year of 15 degrees, it can be calculated that the changes in temperature correspond to changes in electricity consumption of 0.3, which for min-max normalised profile represents 30% of entire 0-1 range.

Next, the average night-time energy consumption in winter as well as the ratio between average night time energy consumption in winter to summer are not as indicative as the two previously mentioned features. The main reason for that is the fact that min-max normalised load profiles have a common range, thus minimising the inter-profile differences in energy consumption, at the benefit of emphasising the potential relations between electricity consumption and temperature as well as minimising the impact of confounding variables such as house size. Nevertheless, it can be seen that even those two features do provide some distinguishing capability among the two categories, judging by their median values.

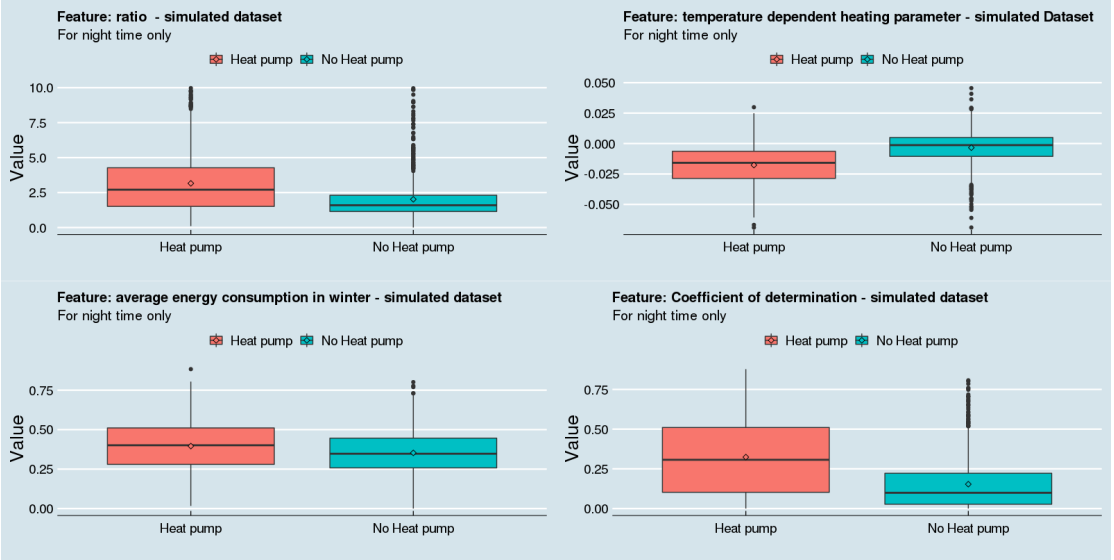

FIGURE 6.8: Conditional box plots of the features for simulated dataset

As far as features obtained form the baseload and simulated dataset are concerned, it can be seen that while the median values reflect what is expected, such as a negative slope, higher average consumption, higher coefficient of determination and higher ratio of energy consumption for owners of heat pumps, the differences among the groups are less pronounced. This indicates that the same set of features provides far better capability of distinguishing owners and non-owners of heat pumps for Alliander dataset, rather than the simulated one.

6.4 Chapter summary

In this chapter the potential feature candidates were elaborated upon along the final list of features and reasoning behind choosing such a set. Furthermore, numerical values of those features have been analysed with emphasis being put on their usefulness in distinguishing between the two categories of owners and non-owners.

Chapter 7

Machine Learning results

7.1 Chapter Introduction

In this chapter the evaluation criteria of machine learning models will be presented alongside the results of the supervised classification problem will be elaborated upon. Results obtained from using logistic regression, decision tree, and various support vector machines, differentiated by the kernel usage, will be discussed with the aim of identifying optimal hyperparameters for each. Once those hyperparameters are identified, the validation set approach is used in order to evaluate how well each of those optimal machine learning models identifies heat pump owners in the simulated dataset.

7.2 Assessment procedure

Several hyperparameters were tuned in the subsequent sections, most important one being class weight. What is meant by class weight is different importance given to observations of one class as opposed to another. Adjusting class weights is crucial since as mentioned in chapter 3 there is significant class imbalance where 90% of all observations in Alliander's dataset, which is used in the first stage to select optimal hyperparameters and later on used as training set for optimal machine learning models, belong to households without heat pump as opposed to 10% with heat pump. Choosing optimal class weight ensure that the model will not classify all observations as negative, seeing the imbalance in the dataset, in order to achieve a high accuracy of 90%. Thus class weights will be swept and evaluation criteria will be assessed for each individual class weights combination. The following procedure was used for creating a matrix of class weights:

1. for i in $\{0, 0.01, 0.02, \ldots, 0.2\}$
   (a) assign value i as weight for the observations representing households without heat pump;
   (b) assign value 1-i as weight for the observations representing households with heat pump;
2. perform 10-fold cross validation with the model of class weights corresponding to ones specified above;
3. record mean, minimum and maximum values of the evaluation metrics;

There are also other hyperparameters which ought to be adjusted. In case of support vector machine models, the values of the cost parameter needs to be adjusted as well. Cost parameter, as mentioned in chapter 3, allows for certain number of misclassifications in case of a non-separable hyperplane. The higher the cost parameter, the smaller the probability of overfitting. Lastly, in case of polynomial kernel, the degree of the polynomial boundary needs to be specified as well. Both cost and degree parameter will be swept in identical fashion to class weights, parameter. The vectors of values for the hyperparameters are:

- degree: is 2,3,4,5;
- cost: 0.001,0.01,0.1,1,5,10,100;
- weights: sequence starting at 0, ending in 0.2 spaced out by 0.01

The procedure for sweeping is the following:

1. for i in vector of degree/cost parameter values:
   (a) assign value i to the degree/cost parameter value of the model;
2. perform 10-fold cross validation with the model of degree/cost parameter corresponding to ones specified above;
3. record mean, minimum and maximum values of the evaluation metrics;

7.3 Evaluation criteria

There are two aspects regarding performance metrics which will be addressed in this section. Firstly, the choice of evaluation metrics will be deliberated by indicating which

ones of those metrics were chosen and why. Next, the benchmark values, that is the target values set by Alliander, stakeholders will be presented alongside explanation on why such values were chosen.

7.3.1 Final choice of evaluation criteria

As far as evaluation criteria are concerned, the following were chosen:

- TNR - this metric represents the percentage of all observations representing households without heat pump that were classified as such. TNR=1 implies all observations without heat pump are classified as non-owners of heat pump.
- TPR - this metric represents the percentage of all observations representing households with heat pump that were classified as such. TPR=1 implies all observations with heat pump are classified as owners of heat pump.
- Precision: details on how this metric is computed are elaborated upon in chapter 4. It helps to understand what percentage of all values identified as positive actually are positive. In this context precision helps evaluates what share of all observations identified as owners of heat pumps actually correspond to owners of heat pumps.
- In the final stages of the analysis of evaluation metrics, the average of TPR, TNR and Precision will be recorded for simplicity, since it provides a single evaluation metric. This average will be referred to as *Mean score*.

Furthermore, apart from the above-mentioned metrics, the AUC will be reported as well. In this context AUC can be see as representing a measure of trade-off between classifying properly all observations corresponding to owners of heat pump at the expense of misclassyfying all observations corresponding households without heat pumps as owners oh heat pump. The higher the AUC the less of a trade-off is present. It should be noted that AUC is not essential in this setting as it is the degree of the trade-off can be evaluated by observing the mean values of TPR and TNR for the same cross-validation instance. Nevertheless, AUC is a popular metric used in this industry thus it is reported here as well.

To understand better what do those evaluation metrics actually mean in the context of this work, let's explore a few hypothetical combinations of TPR, TNR and Precision:

- TNR =1; TPR=1; Precision=1. A perfect classifier which managed to correctly identify all households without heat pump as non-owners of heat pump and all households with heat pump as heat pump owners.

- TNR=1; TPR=0; Precision=NA - none of the values are classified as positive. A poor classifier which classified all observations as negative - non-owners of heat pumps, causing TNR=1, at the expense not being able to detect correctly even a single household with a heat pump. Since there are no positive observations Precision cannot be calculated.
- TNR=0; TPR=1; Precision=share of all positive observations in the entire dataset. A poor classifier which classified all observations as positive - owners of heat pumps, causing TPR=1, at the expense not being able to detect correctly even a single household without a heat pump.

7.3.2 Benchmark values of evaluation criteria

Having established that TNR, TPR and Precision will be used as evaluation criteria, benchmark values, that is the desired minimum mean values for all the three metrics, were chosen. Those values are are 90% for all three: TNR, TPR and Precision.

The reason for choosing such high values stems directly from the way the models developed in this thesis will be used by Alliander. To be more precise:

1. Once the models are developed they will be used to assess which households in the database of 120k load profiles are heat pump owners;
2. these households will be mapped to locations based on their zip codes;
3. based on a separate predictive model, households which will switch to heat pumps in next 1/2/5/10 etc. years will be identified;
4. network congestion will be modelled based on heat pump adaptation rates.

As it can be seen above, a potentially insignificant deviation from the ideal evaluation metric values of triple 1, might cause significant distortions in the *spreidingsmodels*, models aiming to identify which households will switch at what time to heat pumps. An example illustrates well this issue:

- TNR =0.9; TPR=0.9; Precision=0.5. Such a classifier contains significant number of false positives - that is households without heat pumps classified as heat pump owners. Consequence of deploying such model would include overestimation of the effect of heat pump adaptation on the network and subsequent over-investment in network adaptation.

- TNR =0.9; TPR=0.5; Precision=0.9. Such a classifier contains significant number of false negative - that is households with heat pumps classified as heat pump non-owners. Consequence of deploying such model would include underestimation of the effect of heat pump adaptation on the network and thus potential future blackouts. item TNR =0.8; TPR=0.5; Precision=0.7. Such a classifier contains significant number of false negative and false negatives. The effect of deploying such a model would be to identify incorrectly the areas which will be impacted the most by the transition to heat pumps. As a result areas with no need of network upgrades will be upgraded and the opposite would be true for areas with a need for upgrade.

7.4 Logistic regression

As expected, the sweeping of weights has an impact on TPR, TNR and Precision of the LR, as presented on figure 7.1. More detailed overview of the results for LR at the stage of hyperparameter tuning can be found in Appendix A - please refer to figures A.1, A.4, A.3 and A.2.

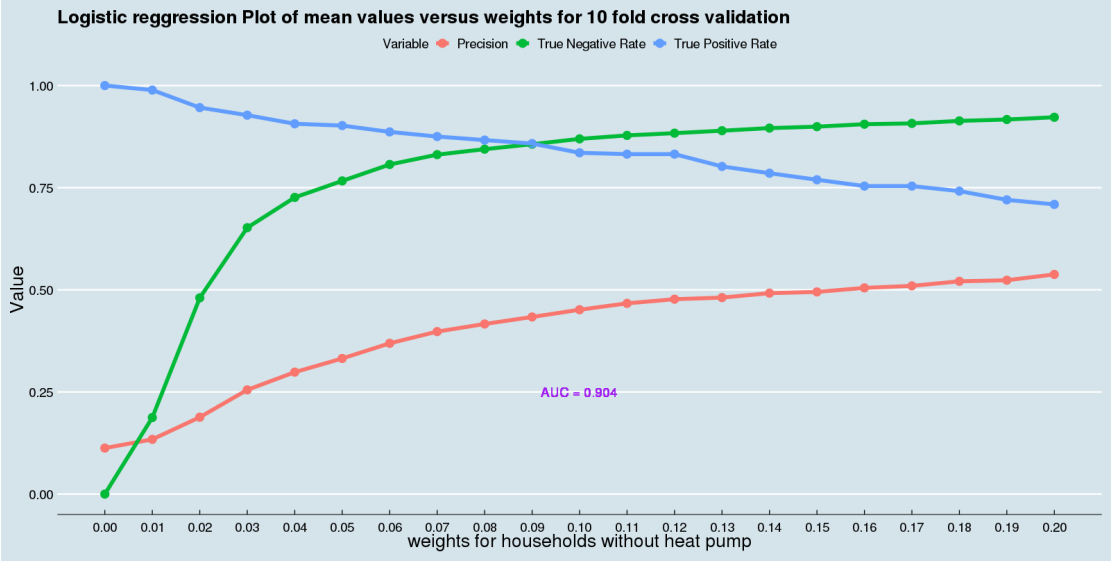

FIGURE 7.1: Plot of evaluation metrics versus class weights for LR

As expected a gradual increase of TNR can be observed as the class weights are increased in favour of observations corresponding to owners of heat pump. As far as the precision metric is concerned, it can be seen that its value steadily rises, from approximately mean 0.15 to just above 0.5, as the weight parameter increases and thus observations of heat pump owners become 'more important'. This however corresponds to a dip in TPR, signalling that increasing share of true positives out of all observations

deemed positive comes at a cost of large number of false negatives, i.e. households with heat pump which are classified as those without one. The AUC, stays relatively constant at 0.9 as the wights are being swept, refer to figure A.1, indicating that the trade-off between separating the two classes is constant.

The decision regarding the optimal choice of hyperparameters for LR model boils down to finding a suitable trade-off between the precision parameter, which is relatively low, and the true positive rate. The optimal trade-off occurs for the class weight of 0.12 when the mean values of AUC, TPR, TNR and Precision are 0.9, 0.9, 0.84 and 0.5 respectively.

7.5 Decision tree

Similar results can be observed for pruned DT, as seen on figure 7.2. The mean value of AUC is just as constant, albeit at a lower level of 0.84, indicating a greater trade-off between TPR and TNR. More detailed overview of the results for DT at the stage of hyperparameter tuning can be found in Appendix A - please refer to figures A.5, A.7, A.6 and A.8.

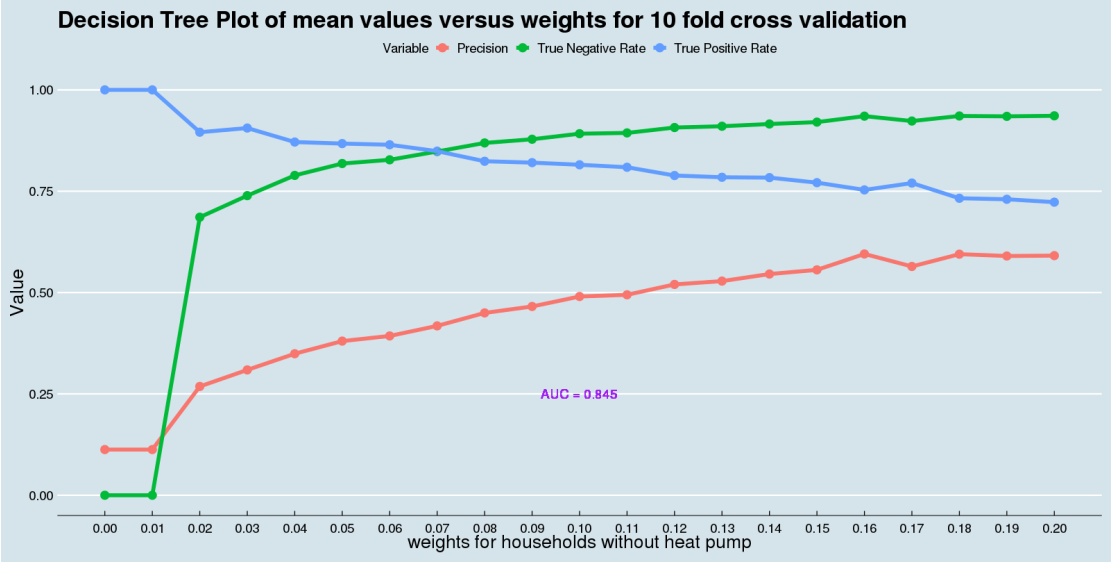

FIGURE 7.2: Plot of evaluation metrics versus class weights for DT

Looking at the TPR, TNR and Precision curves it can be observed that those are less sensitive to the weight parameter as compared to equivalent curves of LR. The Precision curve follows similar trend albeit with more abrupt changes from one weight parameter values to another.

Once again considering that the AUC is stable, please refer to figure A.5, it can be seen the best results can be obtained using weight parameter for households without heat

pump of 0.11, almost identical to the 0.11 for LR case. For the class weight of 0.11 the mean values of AUC, TPR, TNR and Precision are 0.845, 0.8, 0.875 and 0.5 respectively.

7.6 Support Vector Machine results

The SVMs were also used for this classification problem. Considering the tendency of SVMs to be sensitive to hyperparameters as well as taking in consideration the sheer number of hyperparameters to be tuned, each kernel will require separate hyperparameter adjustment. Apart from class weights, cost parameter and polynomial degree, for polynomial kernel SVM only, will be swept.

7.6.1 Kernel: linear (Support Vector Classifier)

First, as in cases of DT and LR, the class weight parameter was swept for the default cost parameter C=1. Subsequently, the cost parameter was swept for class weights parameter fixed at inverse proportionality to the share of labels.

7.6.1.1 Class weights

As it can be seen on figure 7.3 representing the mean evaluation metrics within a 10 fold cross validation while sweeping class weights, there is little change indicating that Support Vector Classifier is largely insensitive to class weights - this occurs for TPR, TNR and Precision which oscillate at around 0.83, 0.83 and 0.45 respectively. Thus the cost parameter will be subsequently swept while keeping the class weight inversely proportional to class weight in the original dataset, that is at 0.11.

More detailed overview of the results for SVM with linear kernel at the stage of class weight tuning can be found in Appendix A - please refer to figures A.9, A.10, A.11 and A.12.

7.6.1.2 Cost parameter

As it can be seen on figure 7.4 representing the mean the evaluation metrics versus the cost parameter, the parameter seems to influence TNR, TPR and Precision by increasing their values with increasing cost, however after cost parameter reaches 1, the values largely converge to 0.82, 0.82 and 0.45 respectively. The values of AUC stands at a high 0.9 for all C values.

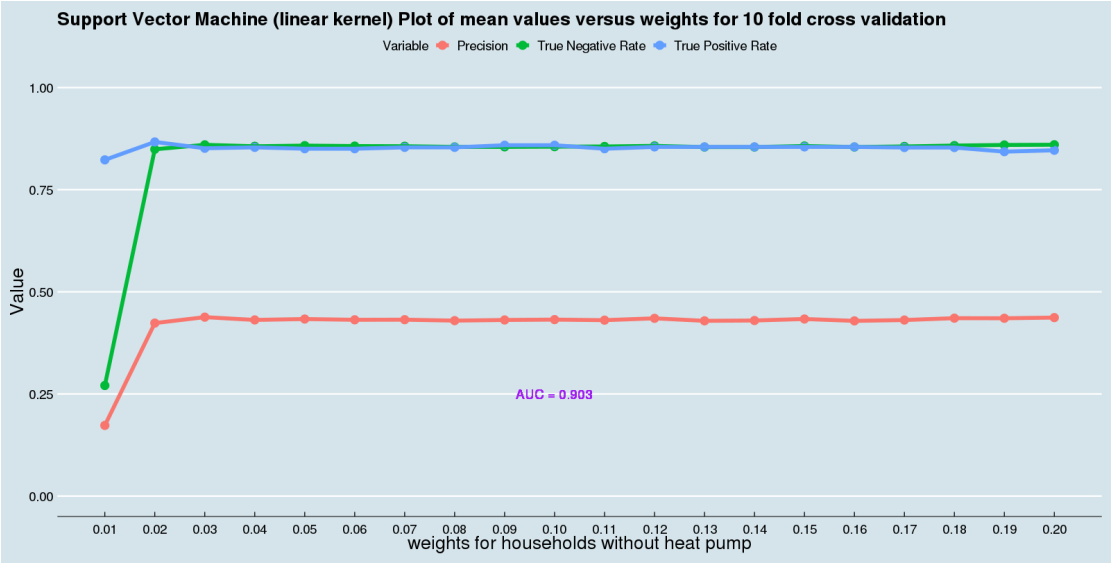

FIGURE 7.3: Plot of evaluation metrics versus class weights for SVM with linear kernel

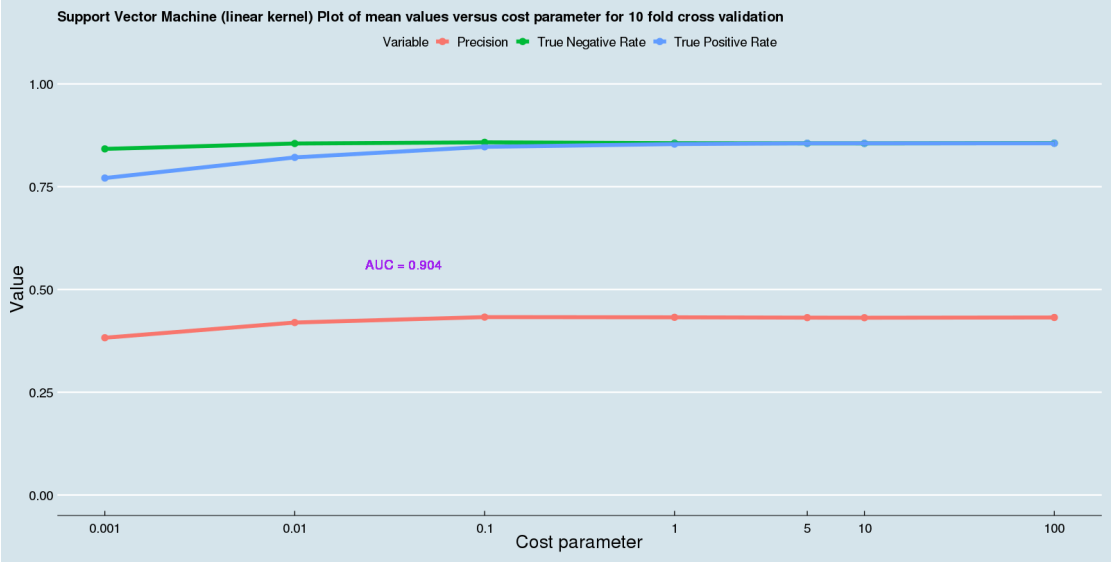

FIGURE 7.4: Plot of evaluation metrics versus cost parameter for SVM with linear kernel

More detailed overview of the results for SVM with linear kernel at the stage of cost parameter tuning can be found in Appendix A - please refer to figures A.13, A.14, A.15 and A.16.

7.6.2 Kernel: radial

First, as in cases of SVM with linear kernel, the class weight parameter was swept for the default cost parameter C=1. Subsequently, the cost parameter was swept for class weights parameter fixed at inverse proportionality to the share of labels.

7.6.2.1 Class weights

Just as in the case of support vector classifier, the class weight parameter did not have a large influence on the AUC which stood at a high 0.92. As it can be seen on 7.5, TPR and TNR slowly increase and decrease respectively with increasing class weight parameter, while precision increases at a faster pace from approximately 0.3 to almost double of that.

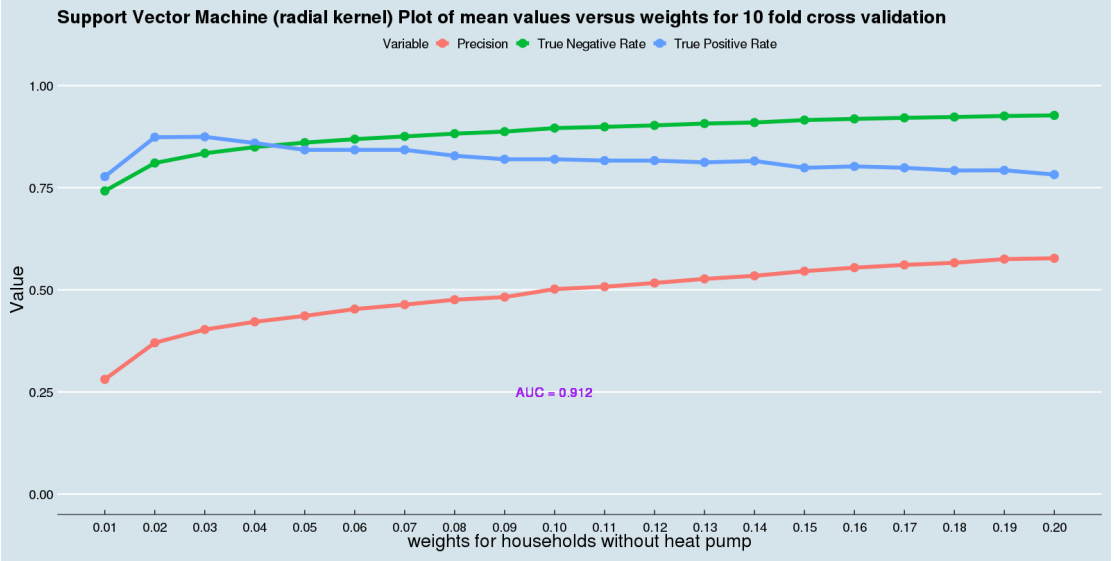

FIGURE 7.5: Plot of evaluation metrics versus class weights for SVM with radial kernel

More detailed overview of the results for SVM with radial kernel at the stage of class weight tuning can be found in Appendix A - please refer to figures A.17, A.18, A.19 and A.20.

7.6.2.2 Cost parameter

As far as the cost parameter is concerned it has a strong influence on Precision, with it staying at 0.15 for C=0.001 and increasing to over 0.5 for C=5 and higher. Similar curve can be observed for TNR with corresponding increases from 0 to 0.88. On the other hand the TPR reaches its highest level as the TNR and Precision metric are at its lowest, which is as expected. As cost parameter increases, TPR stays at around 0.85.

As it can be seen on 7.6, the optimal values of TNR, TPR and Precision is achieved for C=5, when all those values equal to 0.87, 0.85, 0.52 respectively. More detailed overview of the results for SVM with radial kernel at the stage of cost parameter tuning can be found in Appendix A - please refer to figures A.21, A.22, A.23 and A.24.

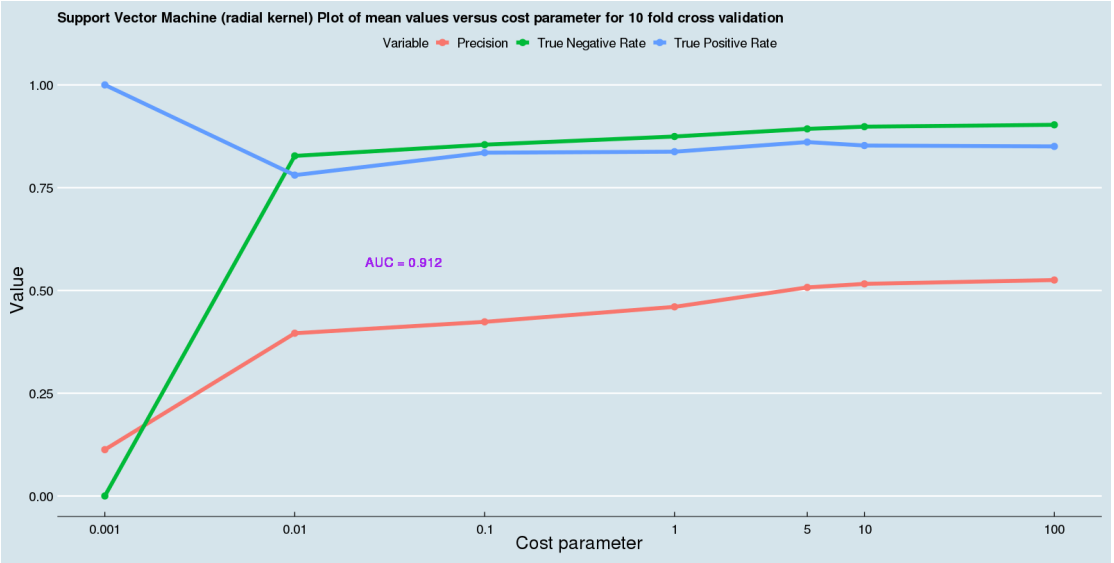

FIGURE 7.6: Plot of evaluation metrics versus cost parameter for SVM with radial kernel

7.6.3 Kernel: polynomial

There are several hyperparameters to be established for a SVM with polynomial kernel. One of them is the degree. Algorithm's performance versus weight, cost and degree parameters has been tested. Polynomial kernel-SVM exhibits AUC equal to 0.9, which is independent of the above-mentioned parameters.

7.6.3.1 Polynomial degree

The degree parameter has an impact on TNR and TPR and Precision as depicted in figures 7.7. The degrees tested ranged from 2 to 5. The reason for not extending the degrees further are two-fold: first the higher the degree the greater the training time. Secondly, degrees above 6 tend to lead to poor results.

In the case of TNR, the values remain stable at the level of around 0.95. As far as TPR is concerned, its value oscillates between 0.75 and 0.6. The Precision evaluation metric on the other hand steadily increases with increasing degree, reaching 0.7 when polynomial degree equals 5. More detailed overview of the results for SVM with polynomial kernel at the stage of polynomial degree tuning can be found in Appendix A - please refer to figures A.33, A.35, A.36 and A.34.

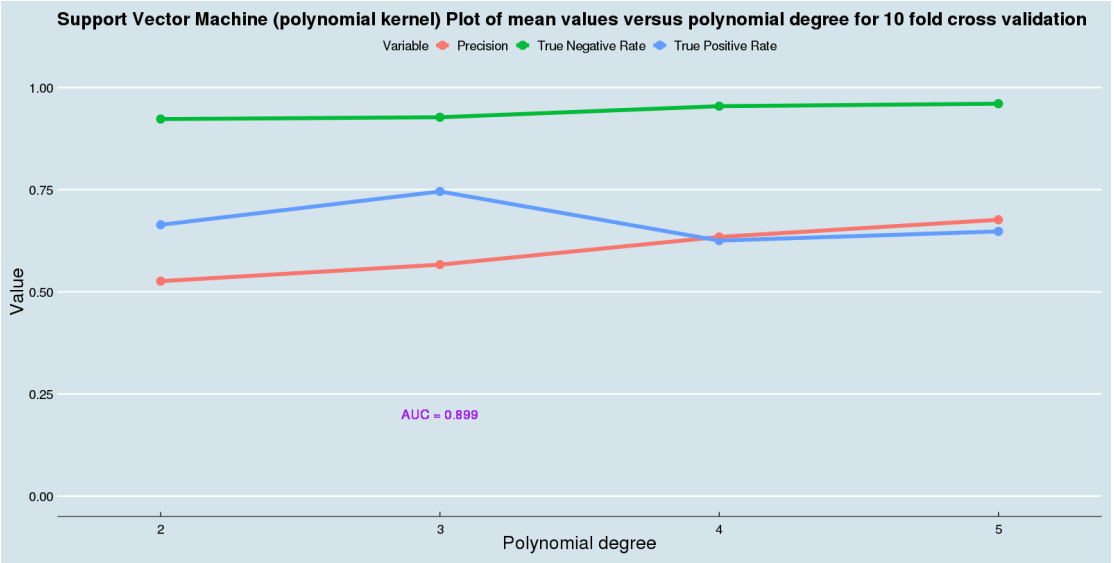

FIGURE 7.7: Plot of evaluation metrics versus polynomial degree for SVM with radial kernel

7.6.3.2 Cost parameter

The cost parameter has an effect on TPR and TNR as can be seen on 7.8. The TNR remains stable at a very high level of 0.95 apart from the cost parameter C=5 for which it takes a dip. Similar trend can be observed for Precision, albeit it stays constant at a lower level of 0.6. The TPR is gradually increasing from around 0.63 to 0.75 achieving its optimal values for C=10.

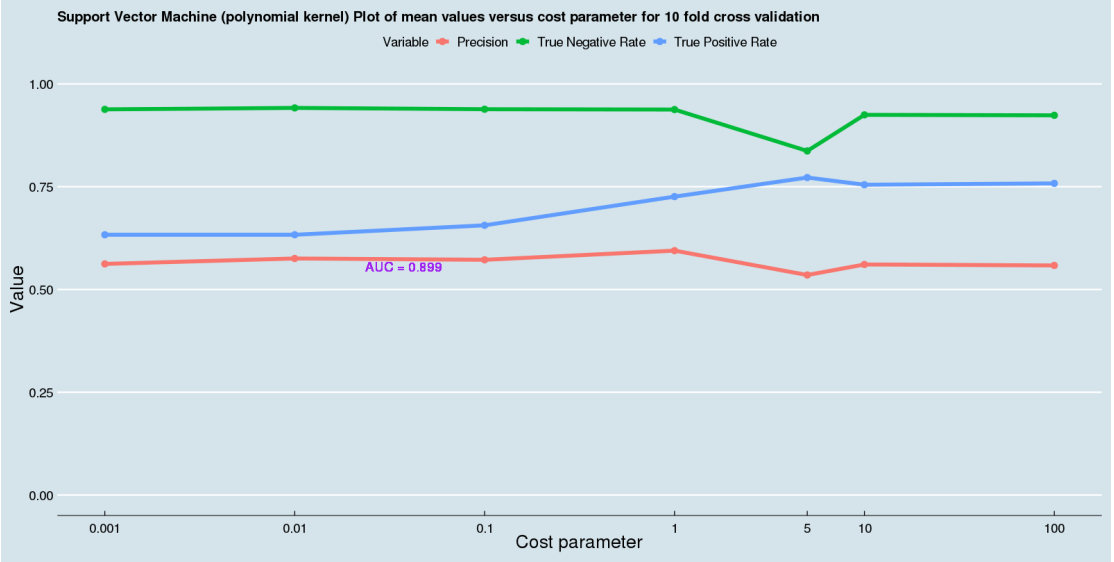

FIGURE 7.8: Plot of evaluation metrics versus cost parameter for SVM with polynomial kernel

More detailed overview of the results for SVM with polynomial kernel at the stage of cost parameter tuning can be found in Appendix A - please refer to figures A.29, A.31, A.32 and A.30.

7.6.3.3 Class weights

The class weights were the final parameter swept across for the three evaluation metrics. TNR and TPR depicted on figure 7.9 are stable for weights greater than 0.05 and reached the levels of 0.9 and 0.75 respectively. The class weight parameter was swept for cost parameter C=10, threshold=0.1, and the polynomial kernel degree of 3 - all of those values were determined beforehand to be optimal.

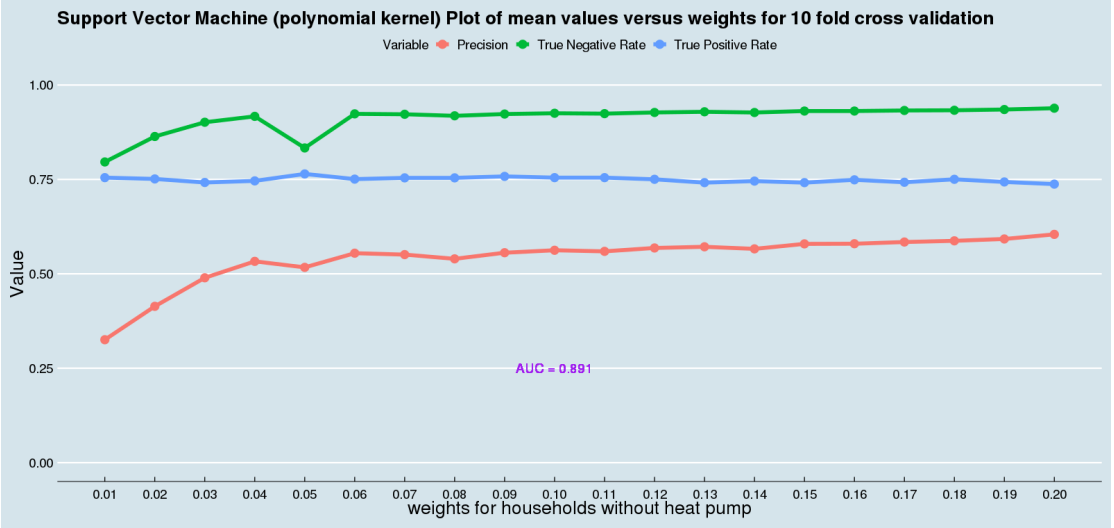

FIGURE 7.9: Plot of evaluation metrics versus class weights parameter for SVM with polynomial kernel

More detailed overview of the results for SVM with polynomial kernel at the stage of class weight tuning can be found in Appendix A - please refer to figures A.25, A.27, A.28 and A.26.

7.7 Models' performance summary with Alliander load profiles

Table 7.1 presents the optimal set of hyperparameters for all of the above-mentioned models. The class weight parameter which is equal to the inversely proportional share of majority labels in the Alliander dataset is optimal for all models. Furthermore, best cost parameter is either 1 or 5 and the best results for polynomial kernel Support Vector Machine are achieved for degree equal to 5.

TABLE 7.1: Table of optimal parameters for each of the tested models

	Class weight	Cost	Degree
Logistic Regression	0.11	X	X
Decision Tree	0.11	X	X
Linear Support Vector Machine	0.11	1	X
Radial Support Vector Machine	0.11	5	X
Polynomial Support Vector Machine	0.11	1	5

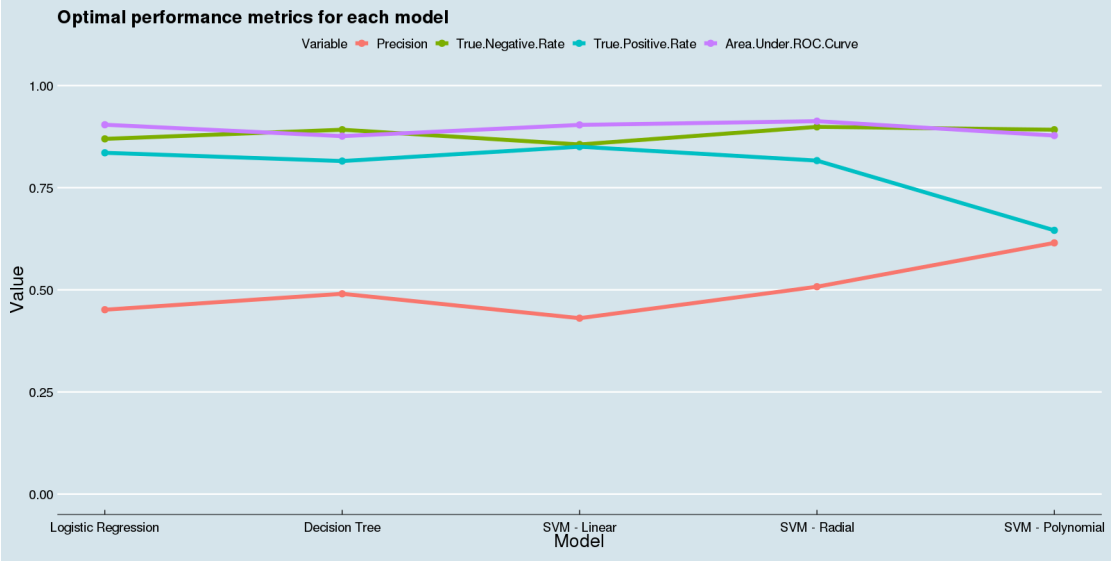

FIGURE 7.10: Plot of mean evaluation metrics within 10-fold cross-validation for each model with optimal hyperparameters

Plot 7.10 contains the values of evaluation metrics for all five models with their best hyperparameters. Separate plot with the average value of the evaluation metrics is seen on 7.11. First it can seen that the mean of all scores changes little from one model to another; what changes are the values of evaluation metrics. Polynomial kernel SVM offers far better performance in terms of Precision, but this comes at a cost of lower TPR. Radial SVM offers performance similar to DT, with the former having slightly higher values for TNR and TPR. The same can be said about the LR, Linear SVM pair. All in all the most balanced performance with the highest mean score is offered by SVM with Radial Kernel.

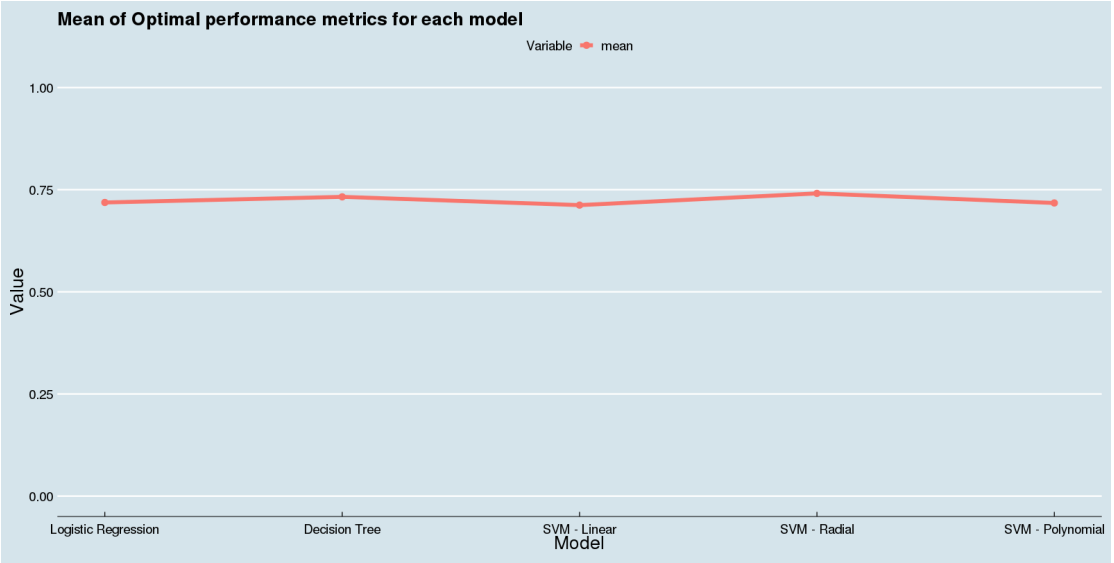

FIGURE 7.11: Plot of mean of evaluation metrics within 10-fold cross-validation for each model with optimal hyperparameters

7.8 Machine Learning performance for simulated dataset

In this subsection performance of machine learning models with optimised hyperparameters with the simulated dataset used as test set, while the feature set of Alliander dataset used as training set will be discussed. In particular, the LR, DT and linear, radial as well as polynomial SVMs with the optimal hyperparameters specified in table 7.1 will be used.

7.8.1 Overview of the results

The final results can be seen on figure 7.12 and the mean score is on figure 7.13. More detailed overview of the results for the machine learning models with tuned hyperparameters can be found in Appendix A - please refer to figures A.37, A.38, A.39, A.40 and A.41.

By comparing figures 7.13 with figure 7.11, it can be seen that the mean score on the simulated dataset does not significantly diverge from the mean metric achieved across all models with the Alliander dataset, around 0.68 versus 0.75 respectively. However, what is different is that result with the simulated dataset exhibit significantly higher precision and significantly lower TPR as compared to the results achieved with Alliander's dataset.

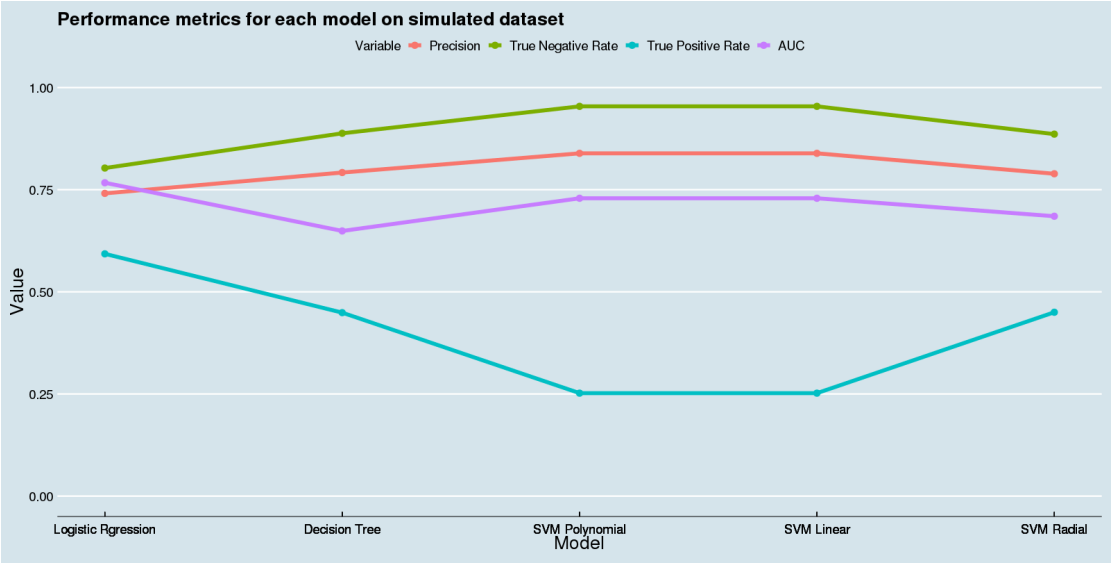

FIGURE 7.12: Plot of mean evaluation metrics with the simulated test set within 10-fold cross-validation for each model with optimal hyperparameters

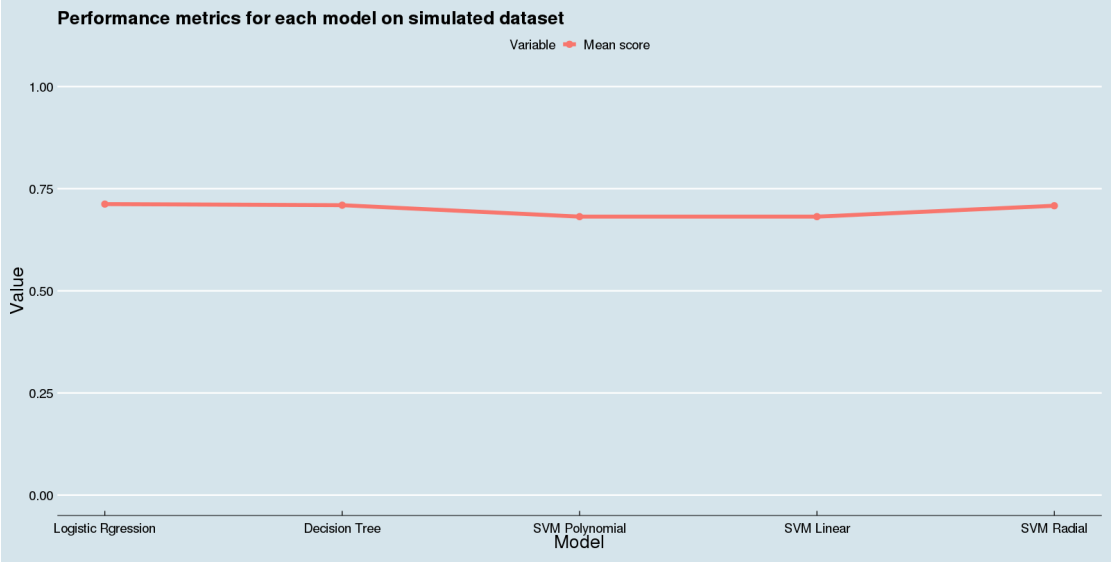

FIGURE 7.13: Plot of mean of evaluation metrics with the simulated test set within 10-fold cross-validation for each model with optimal hyperparameters

7.8.2 Causes of divergence in results

In order to understand what are the causes for the effective 'swapping' of places between precision, it is important to interpret what exactly it means to have high precision and lower TPR in the context of this master thesis.

As a reminder:

- Precision is the ratio between the true positive observations and the sum of all Positive Classifier predictions. In this context it evaluates what percentage of all

customers **classified** as owners of heat pumps, actually have heat pumps.

- TNR is the ratio between the True Negative observations and the sum of all Actual Negative Observations. For the purpose of this thesis it represents the percentage of all **actual** non-owners of heat pumps that were classified as non-owners of heat pump.
- TPR is the ratio between the True Positive observations and the sum of all Actual Positive Observations. For the purpose of this thesis it represents the percentage of all **actual** owners of heat pumps that were classified as owners of heat pump.

7.8.2.1 Analysis of Alliander's dataset results

As mentioned in previous section, results for Alliander's dataset show high TPR and TNR - between 0.8 and 0.95 - and significantly lower Precision - around 0.5. Based on the definitions above it can be concluded that such a case corresponds to large number of false positives. This is because vast majority of owners of heat pumps were classified as such, vast majority of non-owners of heat pumps were classified as such, however, out of all customers classified as owners of heat pumps, at least half did not actually possess one.

It can be thus concluded that there was a significant number of customers that did not actually possess heat pump, whose behaviour was very similar to those customers that did possess one, which explains the appearance false negatives. In order to check for this assumption, a conditional violin plot which is not scaled, that is with the area of the violin corr responding to actual number of observations, was plotted as seen on figure 7.14. While it can be seen that the medians for owners and non owners differ significantly for owners and not of heat pumps, the sheer number of non-owners of heat pumps, 2000 of them in total as compared to 250 of owners, makes it easy for a machine learning model to classify false positives, thus committing type I error.

7.8.2.2 Analysis of simulated dataset results

On the other hand, when looking at the results of machine learning models with simulated dataset being the validation set, it can be noticed that contrary to the results achieved with Alliander's dataset, precision is relatively high and TPR is low, as it can be seen on figure 7.12. This indicates that those classifiers are far more prone to committing type II error, that is false negative, since most of observations classified as owners of heat pumps, are indeed heat pump owners, while a significant number of owners of

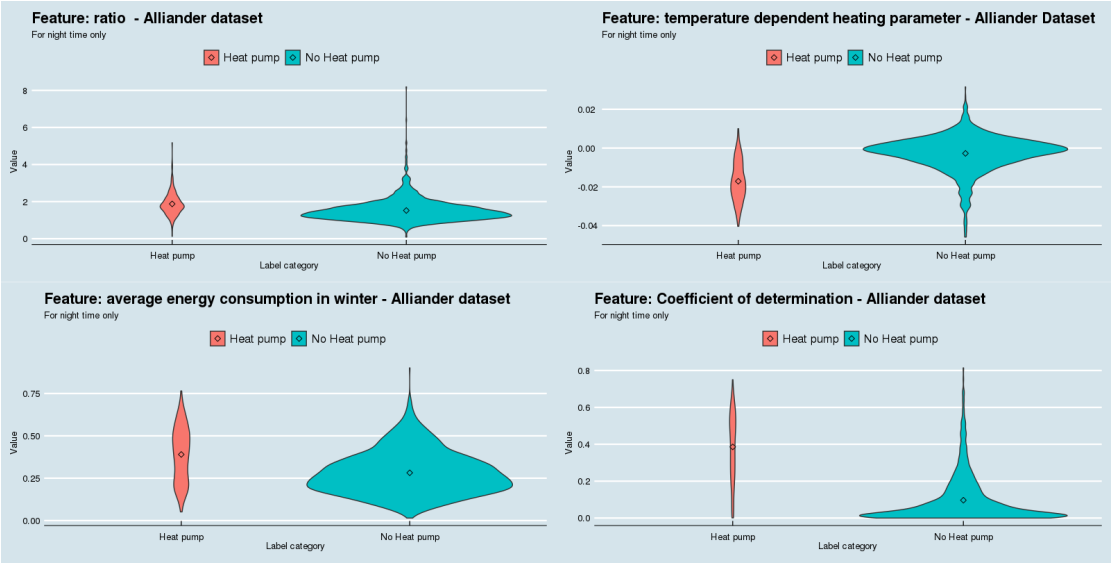

FIGURE 7.14: Conditional violin plot of features extracted from Alliander's dataset

heat pump still remains classified as non-owners - false negative. In other words, there is a large number of heat pump owners in the simulated dataset who exhibit consumption behaviour identical to the behaviour of non-owners. An analysis of feature plot for simulated dataset does confirm this finding, as seen on figure 7.15.

It can be observed that in terms of average energy consumption in winter, there is barely any difference between the owners and non-owners. Furthermore, the differences between the groups temperature dependent heating parameter do not appear large and there is presence of a significant overlap. The only two features which distinguish well two groups are the coefficient of determination and the ratio between energy consumption in winter to summer. This is not surprising seeing how it was the coefficient of determination which was the most indicative feature for Alliander's dataset.

Thus the main question to be asked in order to understand the origin of the divergence in results is:

why features in the simulated dataset are less indicative of owner/non-owner group membership as compared to features in Alliander's dataset?

In order to perform an in-depth analysis, plots of 10th, 90th and 50th(median) percentiles of min-max normalised night-time baseload and heat-pump owner's datasets are are plotted on figure 7.16 and 7.17 respectively. Firstly it can be seen in both cases median and 90th percentile of load profiles tend to increase consumption in winter, albeit the owners of heat pump do it to a greater extent. This explains the relative similarity among the two groups of the temperature heating parameter. While it still holds

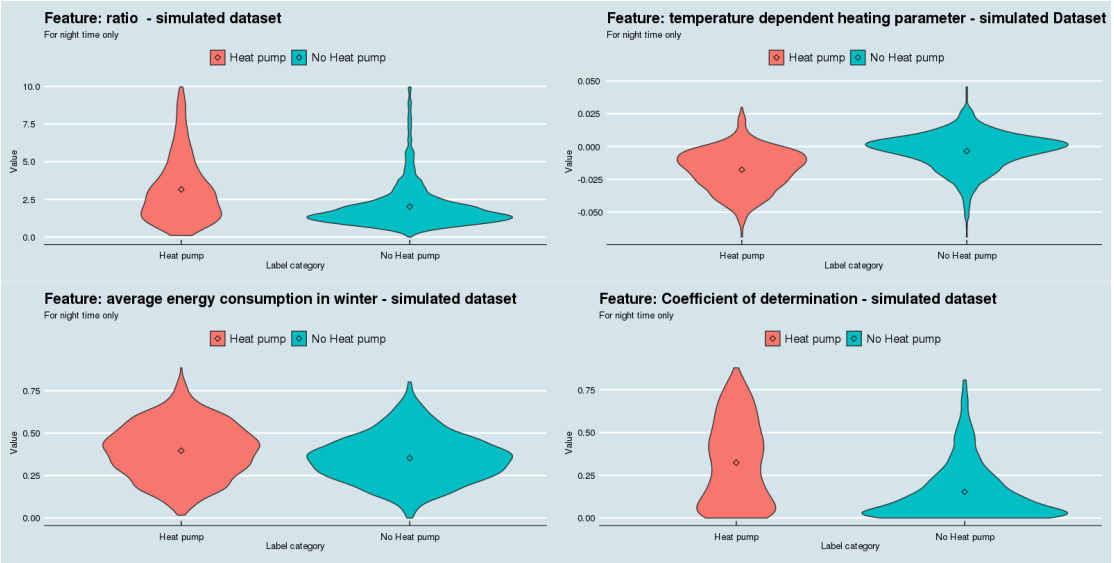

FIGURE 7.15: Conditional violin plot of features extracted from simulated dataset

true that the owners of heat pumps have a more negative slope between the mean daily temperature end daily energy consumption, the median for non-owners is negative as well. Considering however that the rise of energy consumption in winter months is more abrupt and less steady for baseload profiles as compared to the rise for heat pump owners, the coefficient of determination, manages to reflect it and is quite indicative. Such an abrupt rise in energy consumption in baselaod profiles also explains the reason for which the average energy consumption in winter is a poor indicator of membership in the two classes.

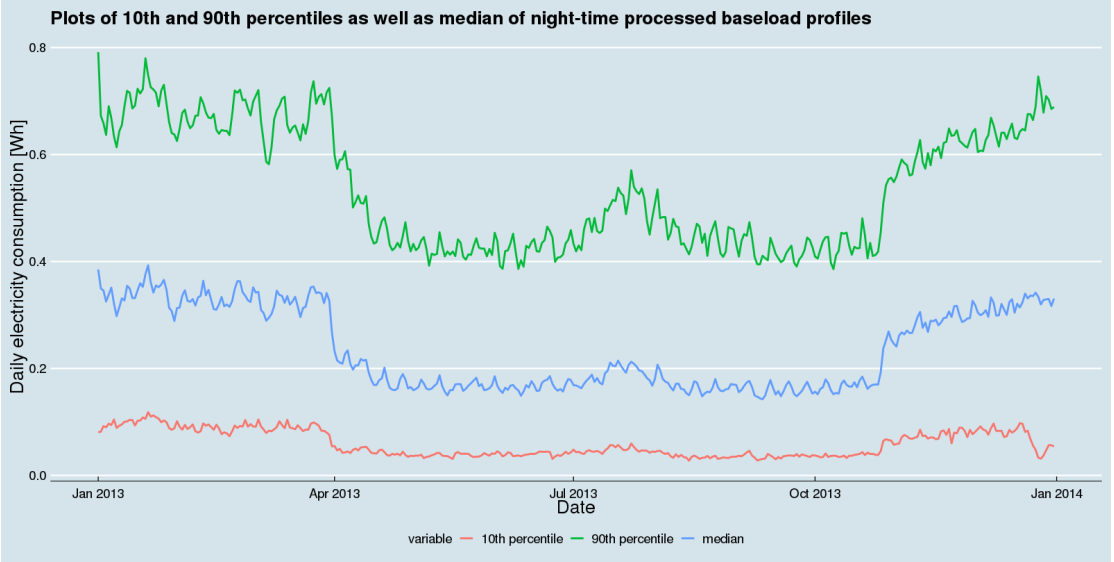

FIGURE 7.16: 90th and 10th percentiles as well as median of electricity consumption of normalised night-time load profiles of baseload profiles

Having proved that the values of features correspond to the overall distribution of load

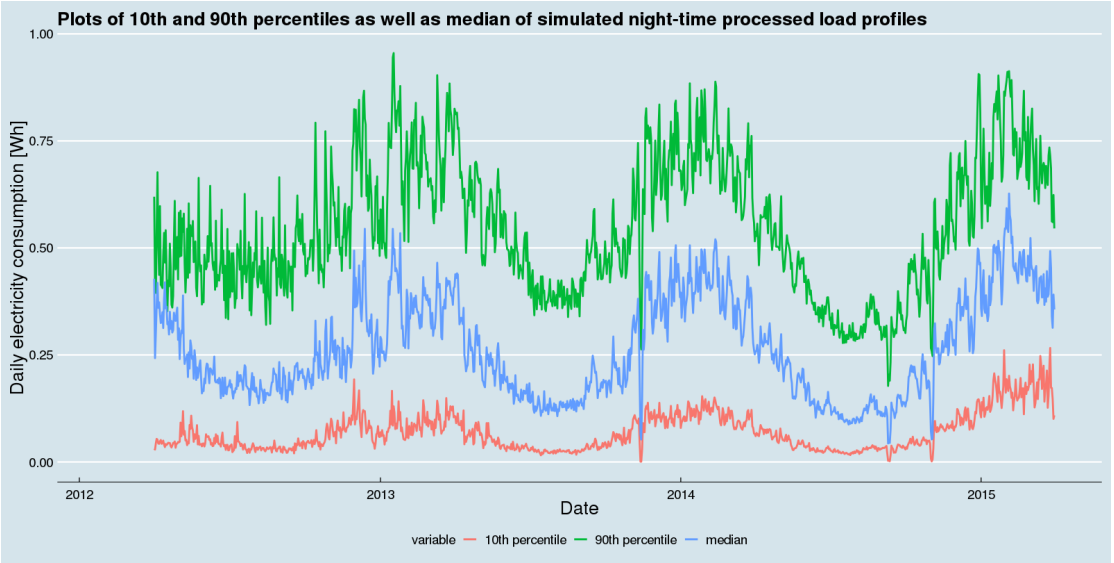

FIGURE 7.17: 90th and 10th percentiles as well as median of electricity consumption of normalised night-time load profiles of simulated heat pump owners

profiles in two groups, it is necessary to ask another question:

why the average energy consumption in winter and the temperature dependent parameter have such different indicative potential for simulated dataset as compared to Alliander's one?

In order to answer the above-mentioned question, a hypothesis was first established:

contrary to heat pump owners in Alliander's dataset, the heat pump owners in simulated dataset, do not use heat pumps as often in the night.

This hypothesis boils down to establishing whether there is a divergence in the heat pump profiles-only consumption between night and daytime. In order to analyse those two periods separately, identical features to the ones used as input to machine learning models but pertaining to daytime usage only where calculated. The conditional violin plots are presented on figure 7.18.

The feature plot proves the hypothesis stated in the paragraph above is correct. The same features as used in input of machine learning models but pertaining to daytime only, that is between 4am and 10pm, are far more indicative of membership of the two groups than the features for the night-time, as can be seen on figure 7.15.

The fact that those identical features are more indicative during daytime as compared to night-time is leads towards conclusion that heat pumps in the simulated dataset are used more frequently during the day as compared to night-time. In order to ascertain with greater confidence this hypothesis 10th, 50th(median) and 90th percentile plots for simulated dataset of owners and non-owners of heat pumps are shown on figures

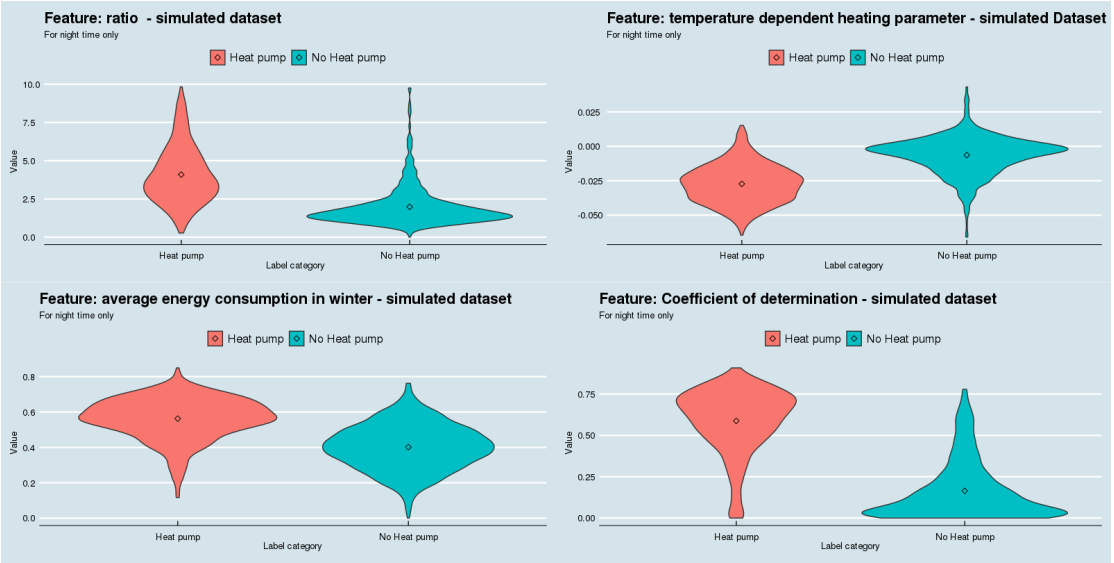

FIGURE 7.18: Conditional violin plot of features extracted from simulated dataset for daytime

7.19 and 7.20 respectively. Those figures do indeed show far greater dependence of load profiles of heat pump owners on outdoors temperature, since gradual increases in winter and subsequent decreases in summer can be observed, while the baseload profiles appear to be more stable.

It is also important to note that it could be that households in the Alliander dataset apart from operating a heat pump also operate water heaters, electrically powered, which draw a constant amount of energy throughout the night thus causing an increase in night-time electricity consumption. It should be thus noted that the statement made in the paragraph above, that the divergence in results between the simulated and Alliander datasets occur due to different heat pump usage patterns, is true only if water heaters are not used at a wide scale - unfortunately there are no sources which could be used to verify this assumption.

Another way of analysing the divergences between the machine learning results is to conclude that the simulation process was incomplete since it assumed that all heat pump owners would be operating non-electrically powered water heaters, while in reality it might be the case that majority of heat pump owners not only install electric space but also water heating.

7.9 Chapter summary

In this chapter the evaluation metrics chosen to assess performance of machine learning models were selected and outline of evaluation procedure was made. Performance of

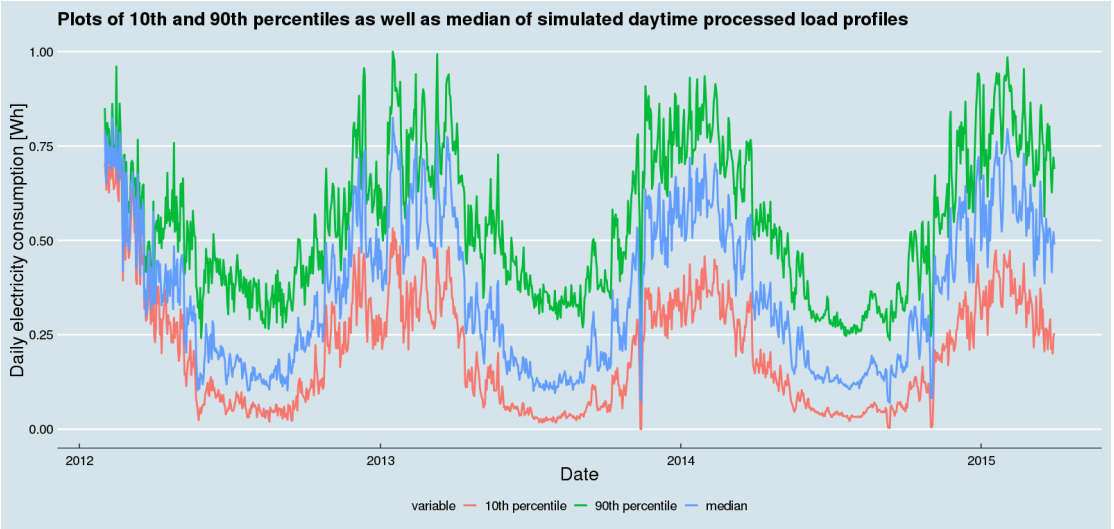

FIGURE 7.19: 90th and 10th percentiles as well as median of electricity consumption of normalised daytime load profiles of simulated heat pump owners

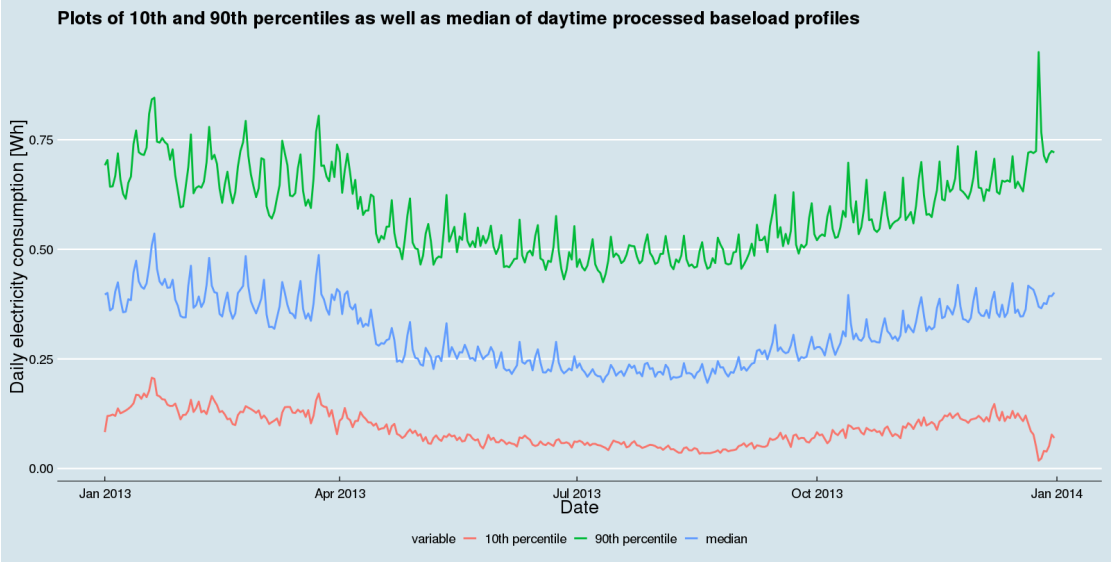

FIGURE 7.20: 90th and 10th percentiles as well as median of electricity consumption of normalised daytime load profiles of simulated heat pump non-owners

machine learning models was presented for two cases: with the test and training set being the feature set based on the dataset provided by Alliander in the first case and in the second case the Alliander's derived feature set being the training set while the dataset of baseload profiles and simulated load profiles of heat pump owners combined being the test set. Due to the differences in performance of the models between the first and second cases, reasoning for those differences was analysed and explanation was provided.

Chapter 8

Conclusions and further work

8.1 Chapter Introduction

In this chapter conclusions in form of answers to the main research (sub)questions are presented and further work is suggested. The latter is divided into two categories: further work which improves this research and one which builds upon this research.

8.2 Answers to research questions

As outlined in chapter 1, this thesis aims to answer a research question, which, in order to systematise approach towards answering that question, was further divided into three sub-questions. Corresponding answers are outlined below.

8.2.1 Answer to main research question

The following was the main research question: Is it possible to accurately predict presence of heat pumps based on load profiles from smart meters?

The answer to the main research question depends on the definition of *accurate*. For Alliander's purposes accurate is defined as achieving mean performance in terms of Precision, TPR and TNR of 90% for all three as explained in chapter 7. Clearly this has not been achieved within this master thesis. It was impossible to achieve this both for Alliander's dataset only as well as for the simulated dataset stage.

8.2.2 Answer to first research sub-question

The following was the first research sub-question: Which classification model performs best for the purpose of heat pumps detection based on load profiles derived from smart meters?

All classifications models perform fairly similarly. Nevertheless, if the mean score is accepted as the ultimate evaluation metric, that is the average of TPR, TNR and Precision, then the SVM with Radial Kernel can be considered as the best performing model, both when used solely on Alliander's dataset and when used with both the Alliander's and simulated dataset. In both cases the mean score was 0.75 and 0.72 respectively.

Nevertheless, other models' performance metrics did not have widely diverging results. The lowest mean score of evaluation metrics on both simulated dataset and Alliander's one being the test set corresponds to linear SVM. Those mean scores are 0.7 and 0.67 respectively.

8.2.3 Answer to second research sub-question

The following was the second research sub-question: Do simulated load profiles of heat pump users based on open-source datasets of heat pump load profiles and load profiles of gas-heated households possess the same characteristics as the real ones?

In this master thesis, the dataset of heat pump only load profiles clearly exhibited differences in terms of patterns of heat pump usage, causing the features calculated from the simulated dataset to be less indicative of heat pump ownership. Nevertheless, these features, which are a reflection of electricity usage, did exhibit characteristics pertaining to those of actual heat pump owners such as negative slope between temperature and electricity consumption, higher ratio of energy consumption in winter to summer etc. To conclude, simulated dataset did possess the same characteristics as the real one, however those differences were less pronounced.

Another reason for which features for simulated load profiles of heat pump owners are less indicative is that most heat pump owners not only use electrically powered space heating but also water heating. Thus during the simulation process an additional component that should have been coupled to baseload and heat pump only load profiles is a load profile of electric water heater. At this stage it is difficult to answer which of the above reasons contributed to the divergence of results.

8.2.4 Answer to third research sub-question

The following was the third research sub-question: How do the machine learning models tested on synthetic load profiles compare to models trained and tested on smart-meter based load profiles?

Machine learning models tested on synthetically created data perform worse than those trained and tested on Alliander's dataset. The difference in performance is not large when the mean score evaluation metric is taken in consideration. On the other hand there is a large divergence in terms of singular evaluation metric scores. Mainly, machine learning models tested on synthetically created data have far higher precision which is accomplished at the expense of having lower TPR, while in case of models trained and tested on Alliander's dataset the opposite is true. This divergence is caused by a different usage pattern of heat pumps in the night-time between the simulated dataset and the one provided by Alliander. A secondary cause for this divergence is the fact that the label distribution among the simulated dataset is equal, 50% for each label category, as opposed to 89%/11% for the Alliander's set.

8.2.5 Additional remarks

While the current machine learning models provide poor precision performance, it the author's belief that as the time passes the precision metric would gradually increase for the same models trained using the same set of features. The reason being, as of right now most heat pump owners are indeed single standing households and thus while min-max normalisation helps to prevent taking the house size in consideration in the first place, it at the same time lowers the precision metric score, since many homes which are not single-standing ones are classified as heat pump owners, even though they are not. This will change however, as increasing number of apartments will start using electricity to power their heating system via shared/district heating system or retrofitted heat pumps. In such a scenario, observations which are false negatives as of the time of this writing, would become heat pump owners, and thus true positives, causing the precision score to increase.

8.3 Further work

Several suggestions regarding potential future work can be made. Those can be divided into two groups:

1. improving the solution developed in this master thesis;
2. using the solutions provided in this master thesis as foundation for another work;

8.3.1 Improving this research

As far as suggestions for improvements of this thesis are concerned several proposals can be made:

- Only four features were used. The main reason for such small number is the fact it was required for those features to be empirical, that is one's which are easy to interpret. As an example:
  - a feature pertaining to daily electricity consumption during a specific day in a year is not empirical, as there might be no particular reason for which that particular day was chosen over the other. Furthermore, such feature might be useful for that particular dataset, but it would likely outrun its usefulness once the model was deployed on an updated dataset.
  - a feature such as coefficient of determination is interpretable and empirical as its use was justified by the expected electricity usage patterns of heat pump owners;

  Nevertheless, several other features could have been used. These include standard deviation of each min-max normalised profile (or any other measure of variance), auto-correlation value, partial auto-correlation values, mean value for each profile;
- Furthermore, rather than normalising each load profile using min-max normalisation process, each profile could have been simply divided by its mean. This would ensure that mean of each profile was 1, while at the same time not forcing each load profile into identical range, which could have made it more difficult to identify heat pump owners;
- More models could have been used. The most popular alternatives to the models presented in this thesis are random forest, Naive Bayes, AdaBoost and Extreme Gradient Boost;
- The data could have been simulated with a dataset in which a significant share of heat pumps is used in the night contrary to the current situation. Thus a new heat pump dataset could have been used for simulation purposes.

- The accuracy and of the research results would be significantly higher if specific location of each load profile, both in Alliander's set as well as simulated one, was known. If this was the case exact temperature recording could have been extracted. This would increase reliability of the results, since no assumptions regarding geographic spread of the load profiles would have to be made.

8.3.2 Building upon this research

Once heat pump owners have been identified based on the pattern in electricity consumption, the next step could be to attempt to identify them based on voltage profiles only. There are several advantages of using voltage profiles rather than load profiles. The first advantage is that voltage profiles are not deemed privacy-sensitive and thus DNOs have easier access to them as compared to load profiles. Next, load profiles under disposal of Alliander are frequently not the most recent ones. On the other hand most recent voltage profiles are, thus increasing the accuracy of results.

Subsequently, since the owners of heat pumps have already been identified, it would be just as useful to identify the type of heat pumps they utilise, seeing that different electricity consumption patterns might be associated with each type. Furthermore, if data on ownership of heat pumps was coupled with socio-demographic factors such as family size, earnings, etc. a prediction for the clients that are the most likely to purchase heat pumps in the near future can be made. In addition to it, if house size was known, not only in terms of inhabitants but also the overall area, users with disproportionately high usage could be identified and notified, giving them a chance to lower their energy bill in the process.

As of the writing of this thesis none of the models presented in this report has actually been deployed on the large dataset of 120k load profiles at disposal of Alliander. Deploying it would allow Alliander to gain some insight into their client base, albeit with a certain confidence interval.

Lastly, having access to some 881 load profiles of heat pump users from simulated dataset and 250 load profiles of heat pump users from load profile data, the electricity consumption versus temperature curve could be constructed in order to better understand the relation between the independent parameter (outdoor temperature) and dependent parameter (electricity consumption). Such a curve could be then coupled with weather forecasts and information on the number of heat pump users within a given area to assess if the infrastructure is properly prepared for the upcoming surge in electricity consumption due to lower temperature etc.

Appendix A

Appendix A: Detailed graphs of machine learning results

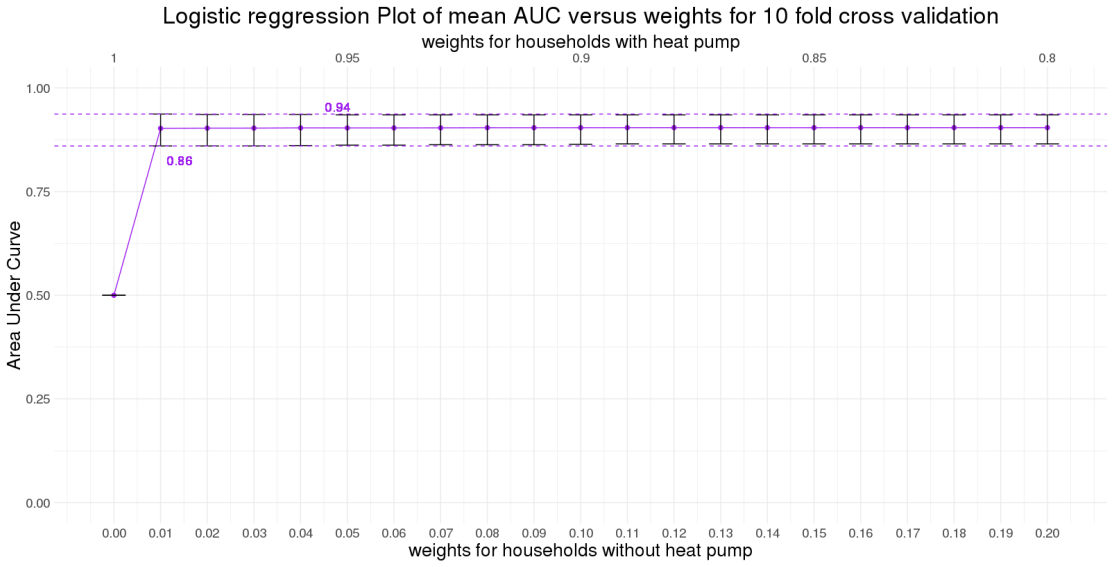

FIGURE A.1: The plot of mean Area under Receiver operating characteristic curve for Logistic Regression within a 10 fold cross-validation for different weight parameters.

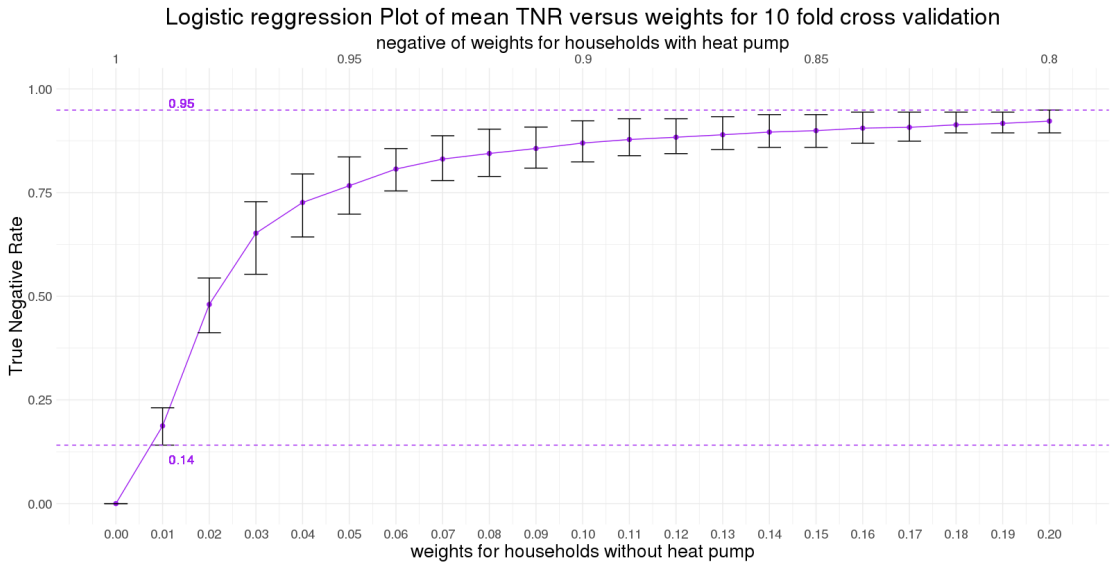

FIGURE A.2: The plot of mean True negative rate for Logistic Regression within a 10 fold cross-validation for different weight parameters.

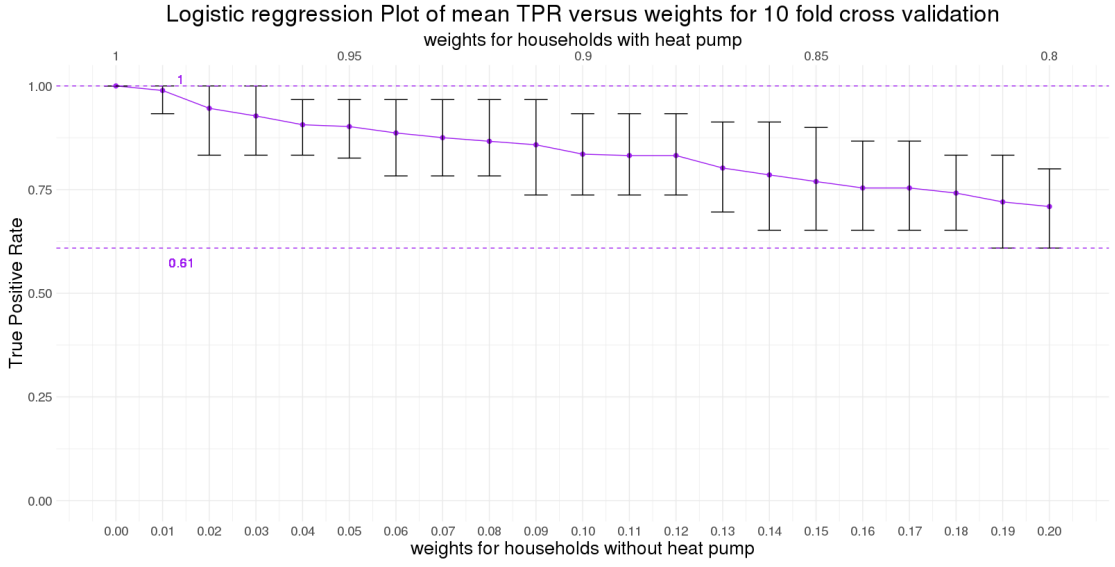

FIGURE A.3: The plot of mean True positive rate for Logistic Regression within a 10 fold cross-validation for different weight parameters.

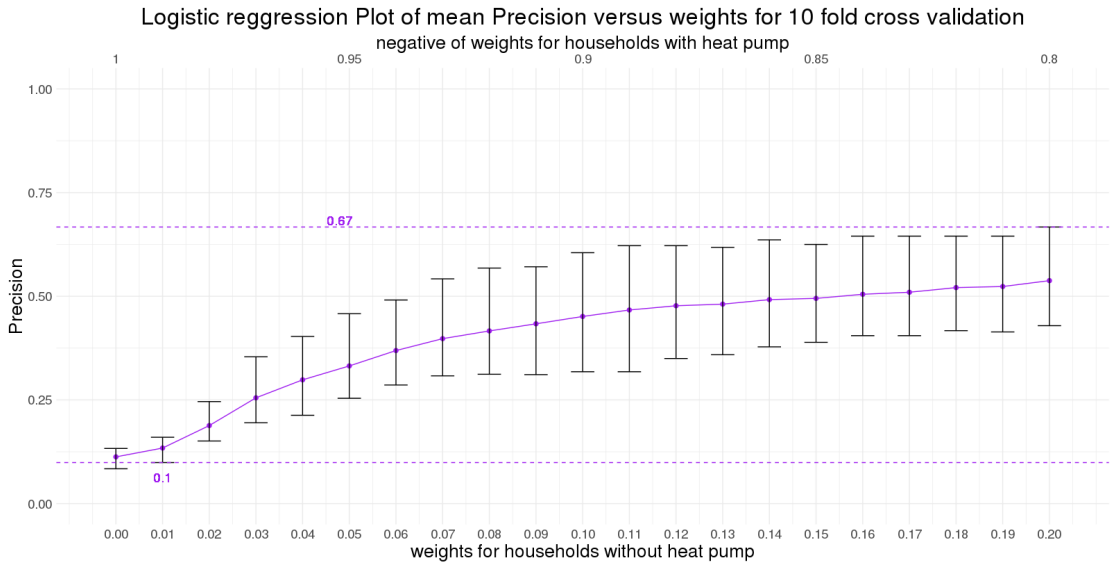

FIGURE A.4: The plot of mean precision for Logistic Regression within a 10 fold cross-validation for different weight parameters.

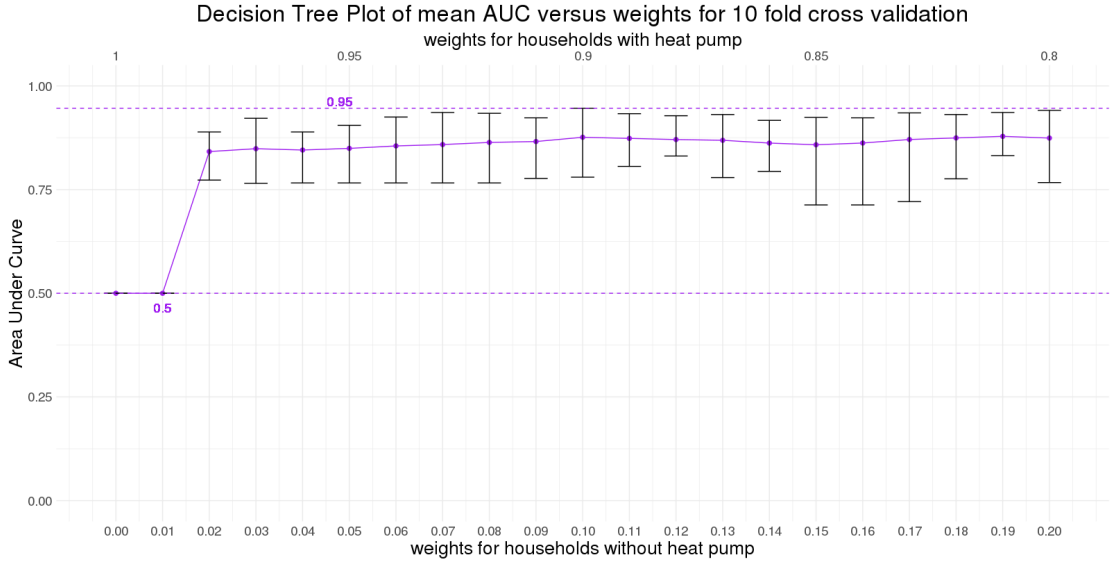

FIGURE A.5: The plot of mean Area Under Receiver Operating Characteristic for Decision tree within a 10 fold cross-validation for different weight parameters.

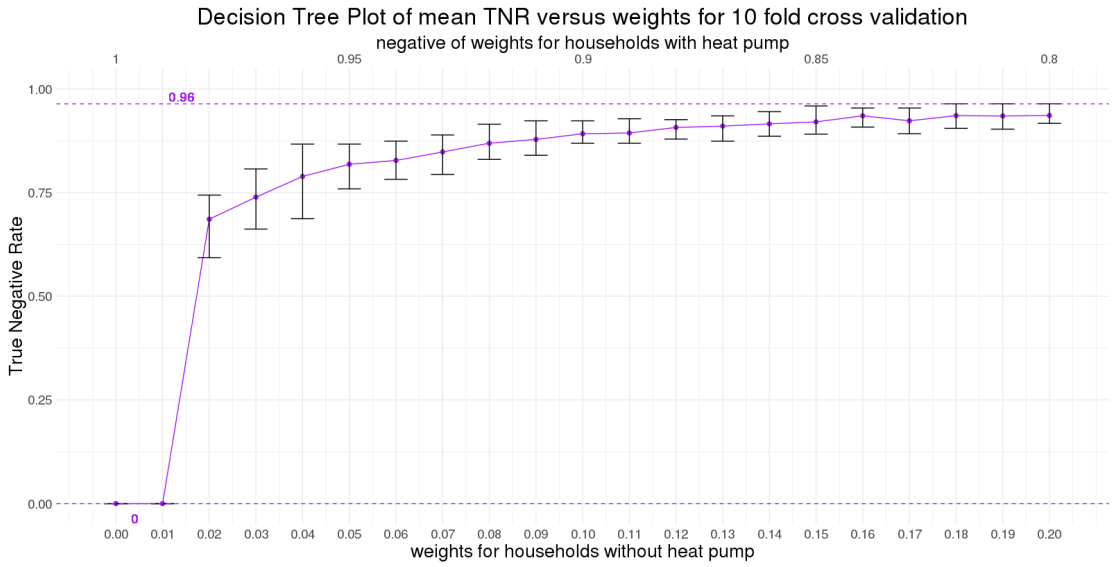

FIGURE A.6: The plot of mean True negative rate for Decision tree within a 10 fold cross-validation for different weight parameters.

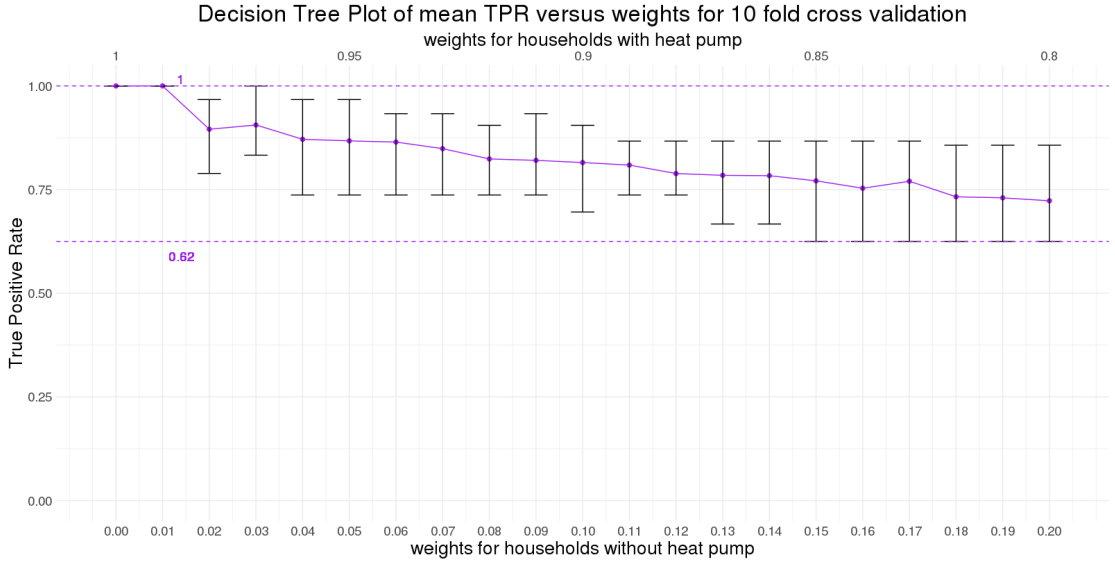

FIGURE A.7: The plot of mean True positive rate for Decision tree within a 10 fold cross-validation for different weight parameters.

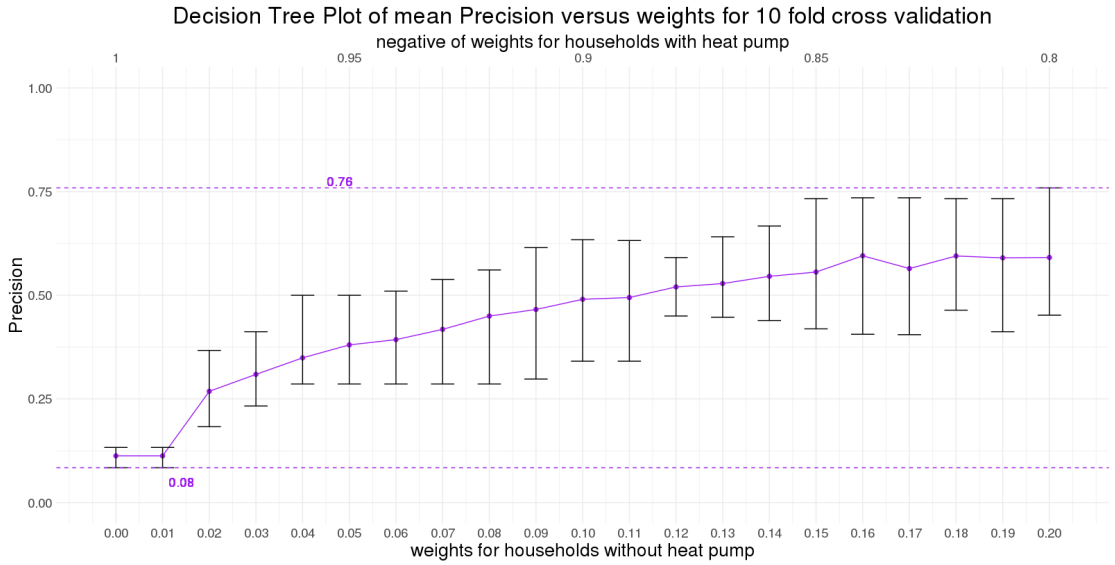

FIGURE A.8: The plot of mean Precision for Decision tree within a 10 fold cross-validation for different weight parameters.

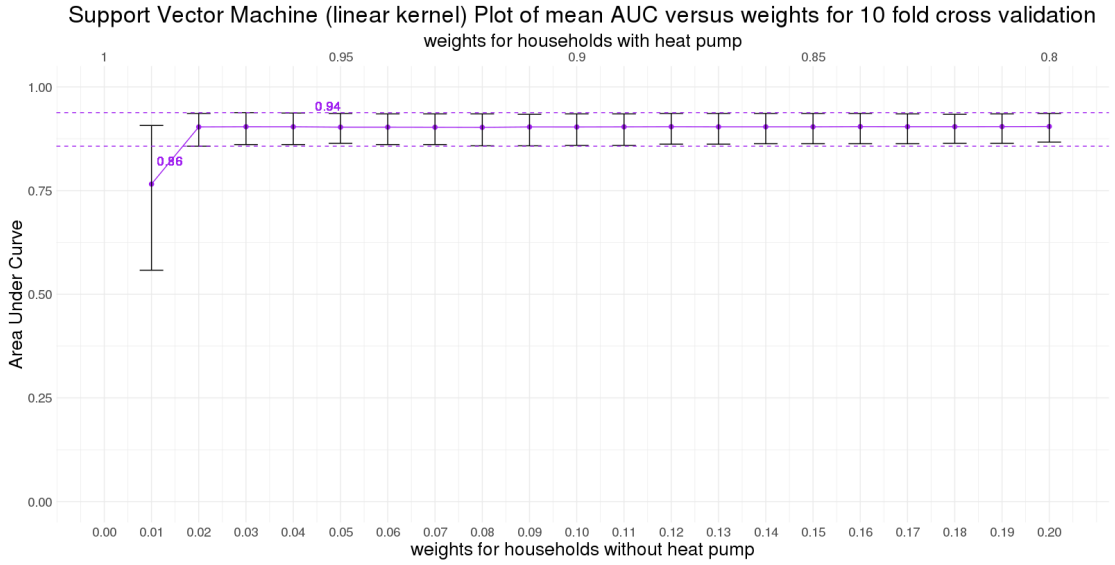

FIGURE A.9: The plot of mean Area Under Receiver Operating Characteristic for Support Vector Classifier within a 10 fold cross-validation for different class weight parameters.

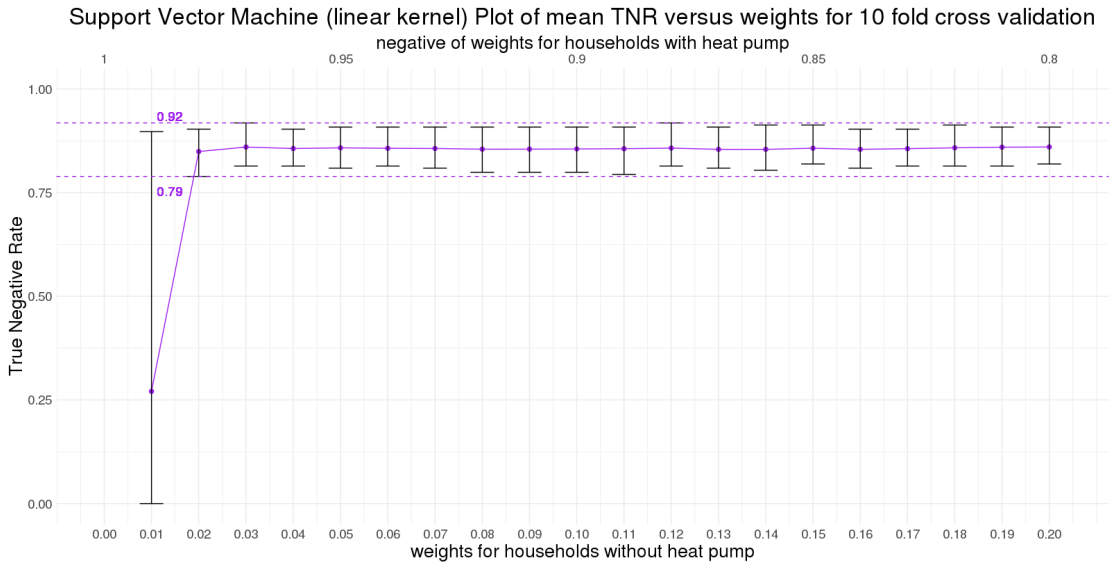

FIGURE A.10: The plot of mean True Negative Rate for Support Vector Classifier within a 10 fold cross-validation for different class weight parameters.

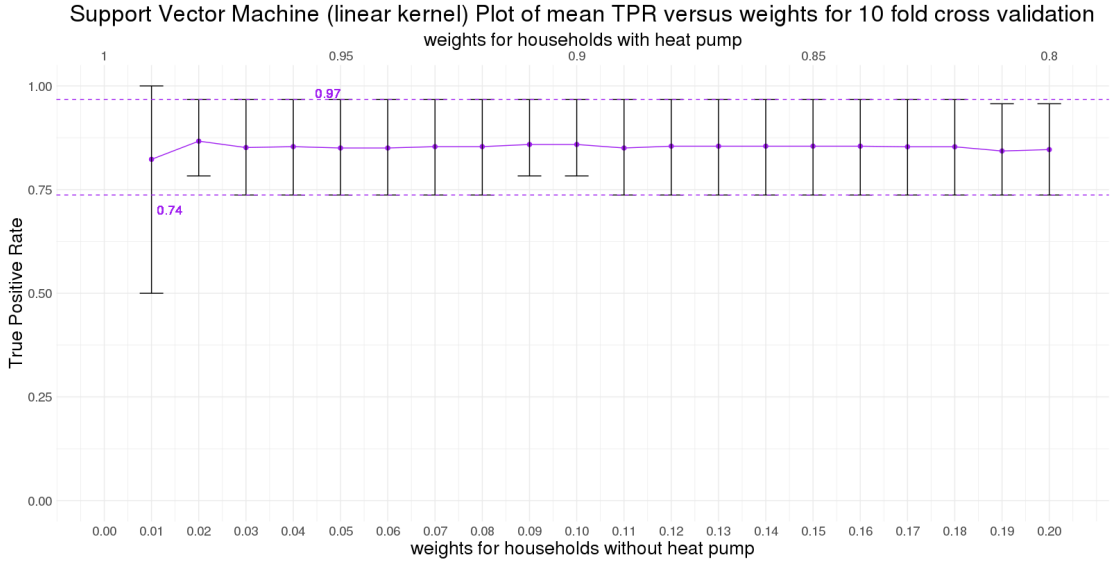

FIGURE A.11: The plot of mean True Positive Rate for Support Vector Classifier within a 10 fold cross-validation for different class weight parameters.

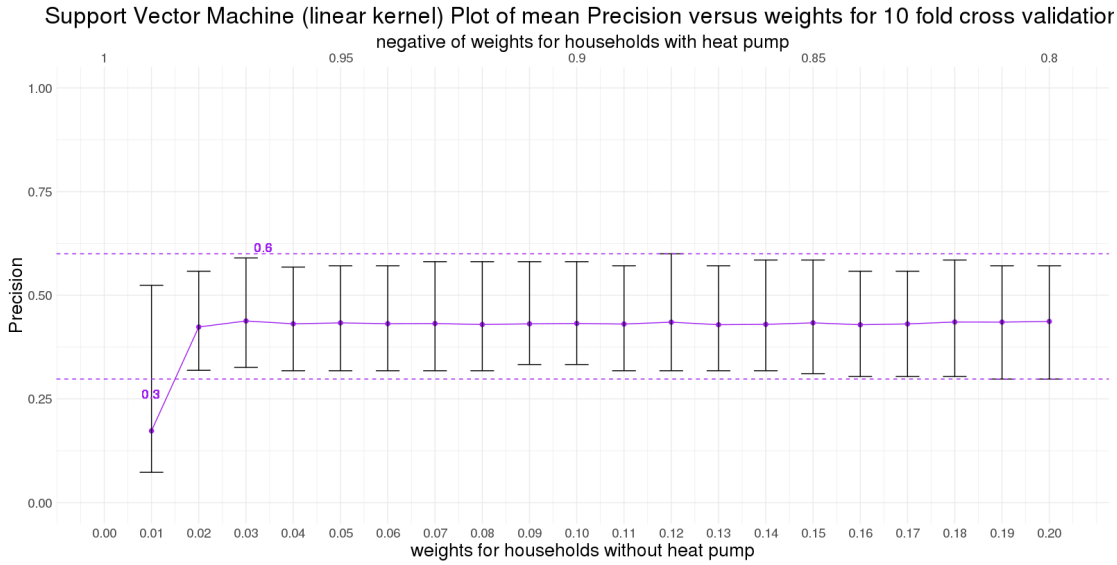

FIGURE A.12: The plot of mean Precision for Support Vector Classifier within a 10 fold cross-validation for different class weight parameters.

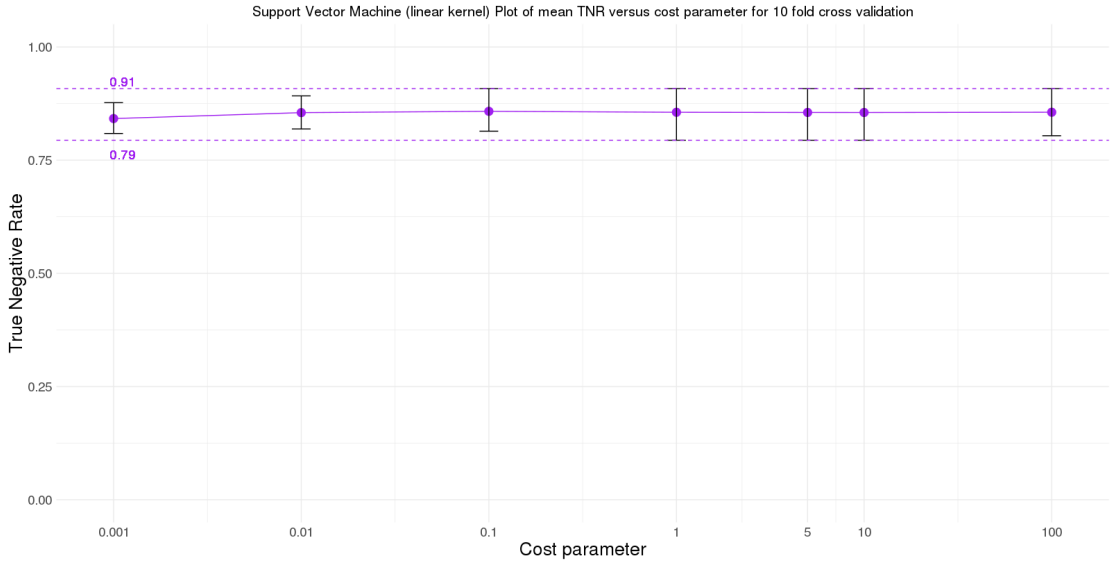

FIGURE A.13: The plot of mean Area Under Receiver Operating Characteristic for Support Vector Classifier within a 10 fold cross-validation for different cost parameters.

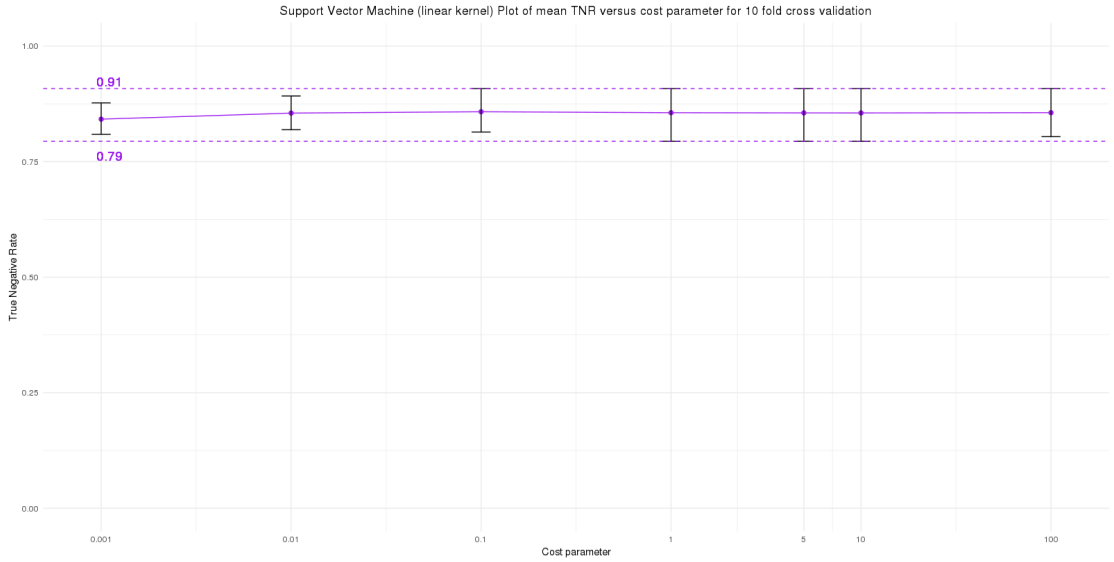

FIGURE A.14: The plot of mean True Negative Rate for Support Vector Classifier within a 10 fold cross-validation for different cost parameters.

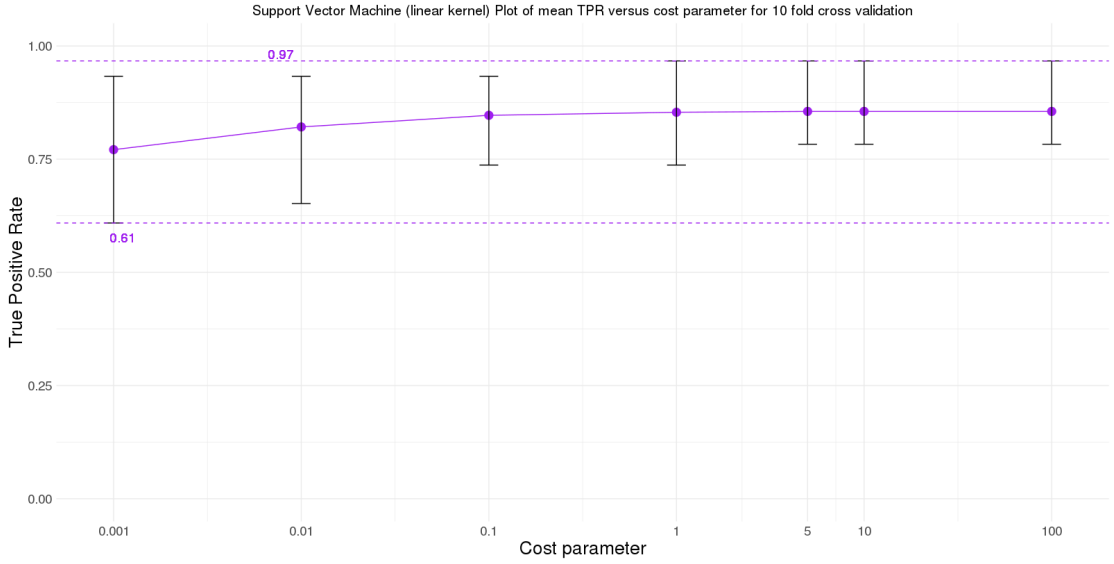

FIGURE A.15: The plot of mean True Positive Rate for Support Vector Classifier within a 10 fold cross-validation for different cost parameters.

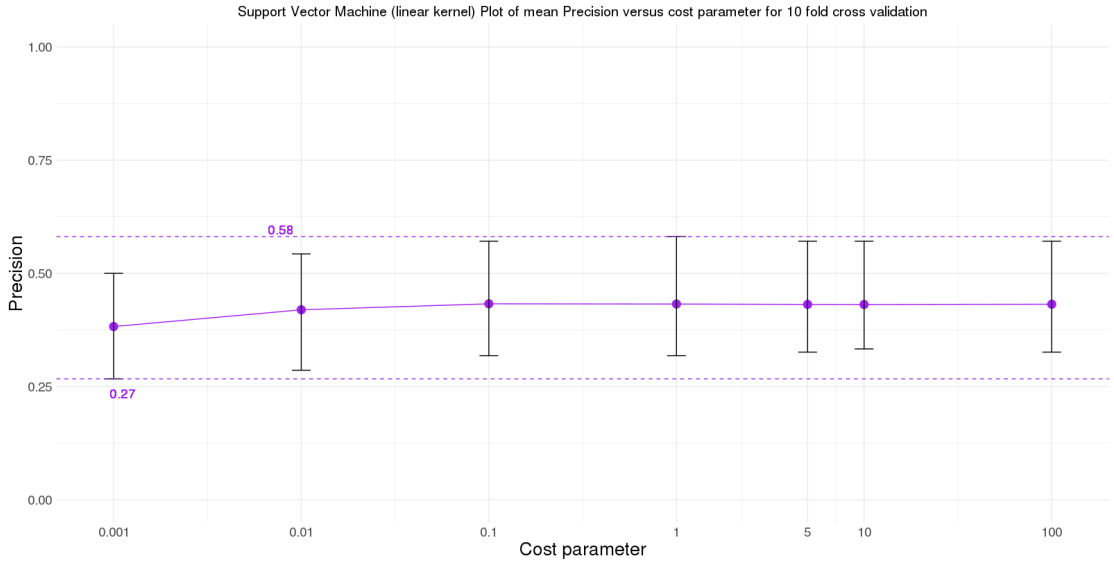

FIGURE A.16: The plot of mean Precision for Support Vector Classifier within a 10 fold cross-validation for different cost parameters.

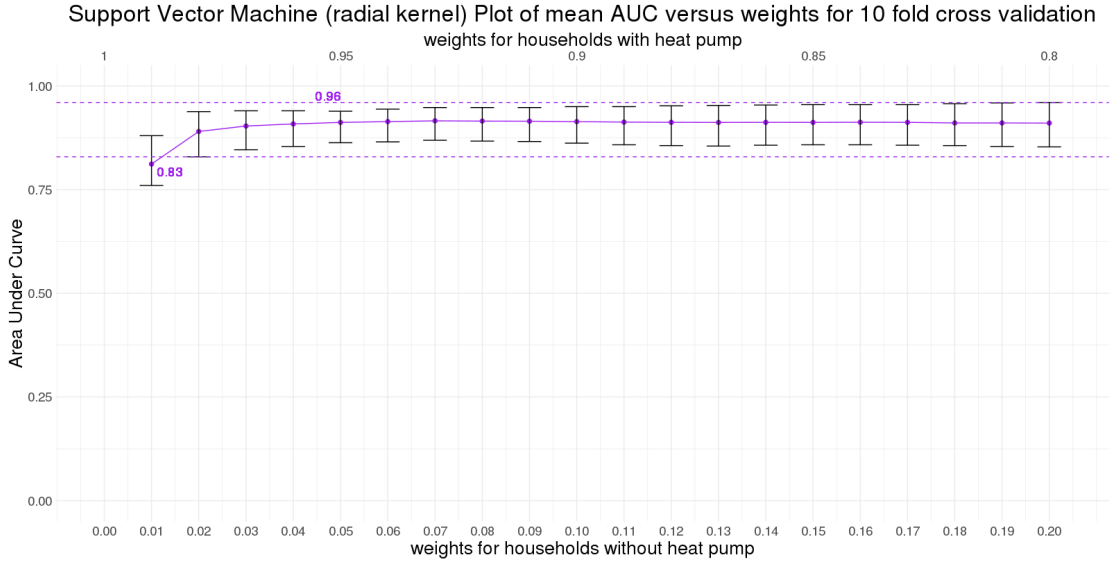

FIGURE A.17: The plot of mean Area Under ROC Curve for Support Vector Machine with a radial kernel within a 10 fold cross-validation for different weight parameters.

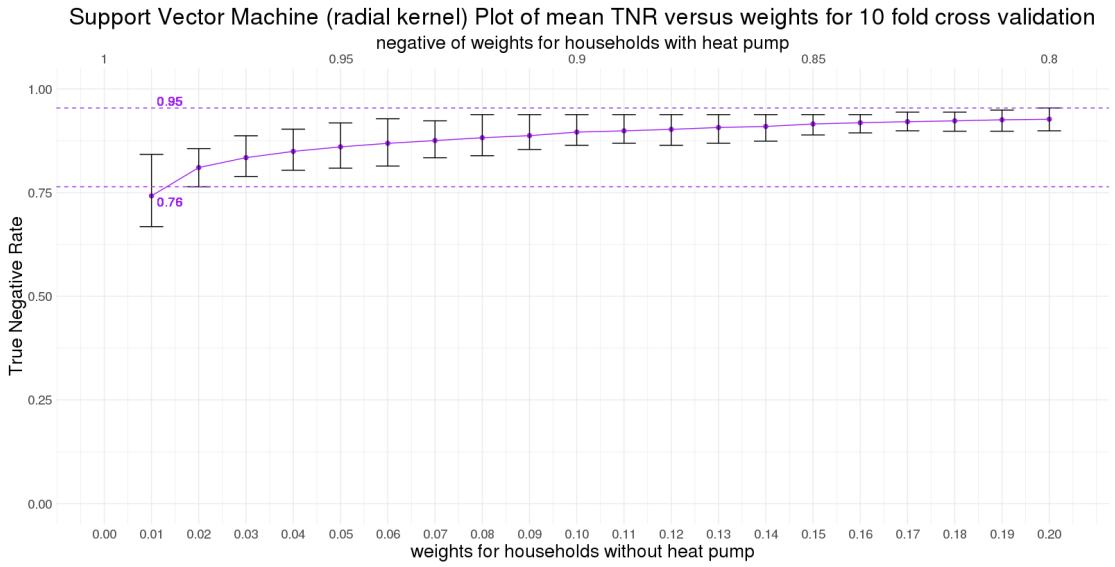

FIGURE A.18: The plot of mean True Negative Rate for Support Vector Machine with a radial kernel within a 10 fold cross-validation for different weight parameters.

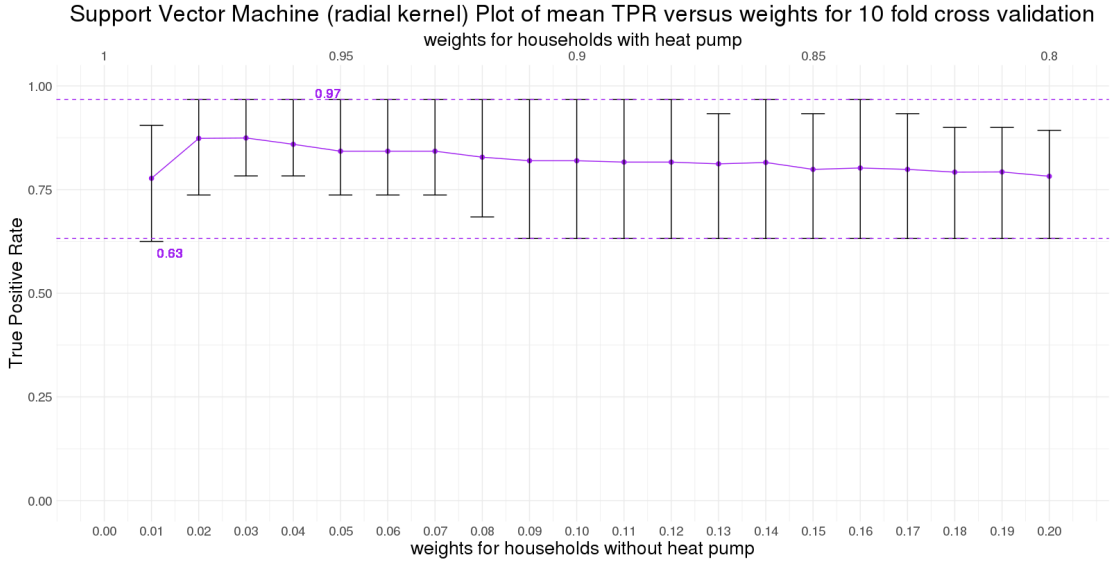

FIGURE A.19: The plot of mean True Positive Rate for Support Vector Machine with a radial kernel within a 10 fold cross-validation for different weight parameters.

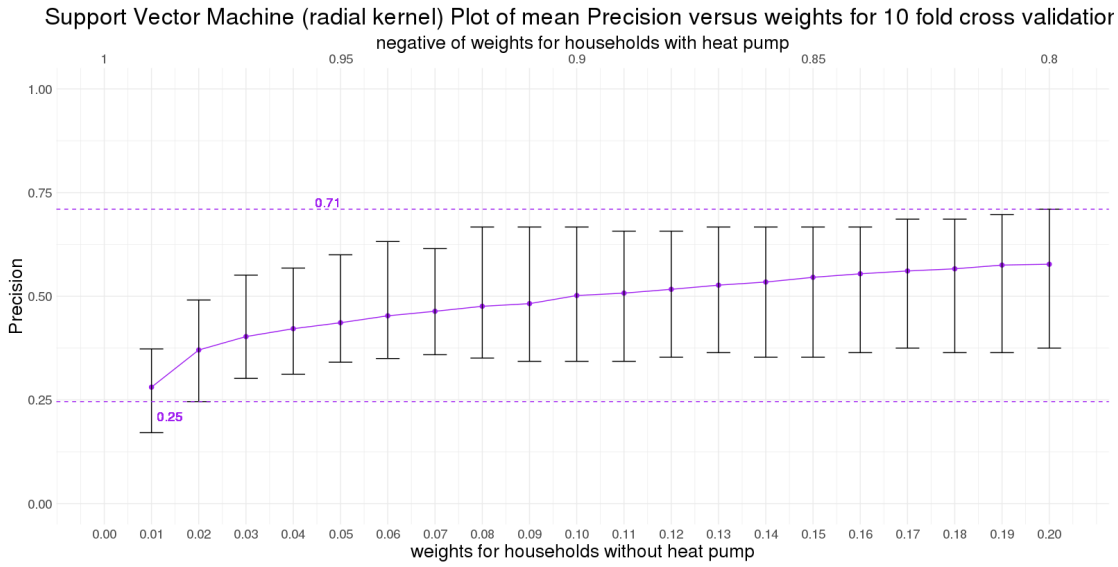

FIGURE A.20: The plot of mean Precision for Support Vector Machine with a radial kernel within a 10 fold cross-validation for different weight parameters.

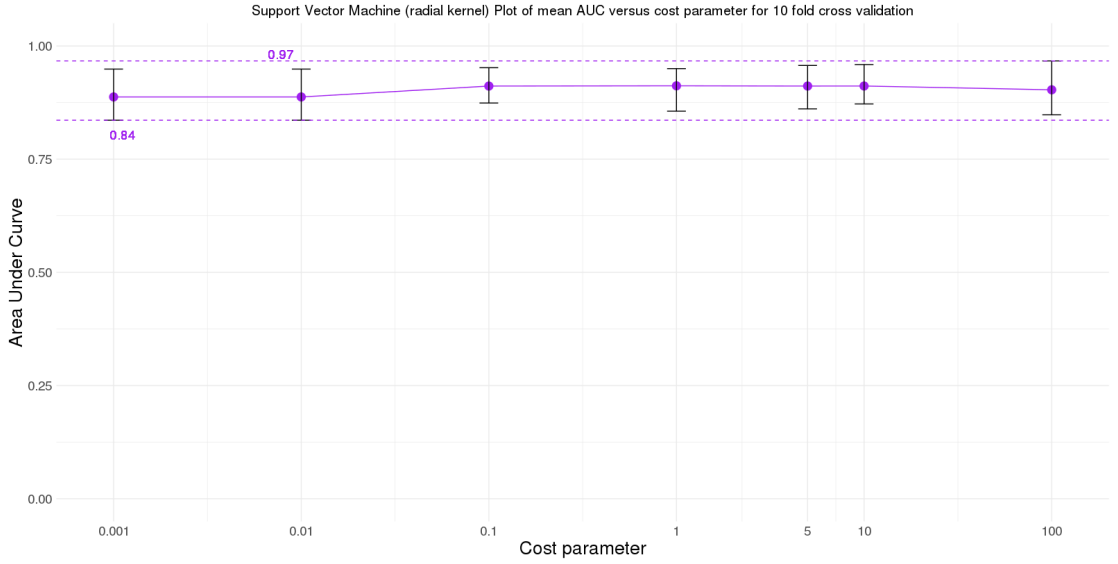

FIGURE A.21: The plot of mean Area Under ROC Curve for Support Vector Machine with a radial kernel within a 10 fold cross-validation for different cost parameters.

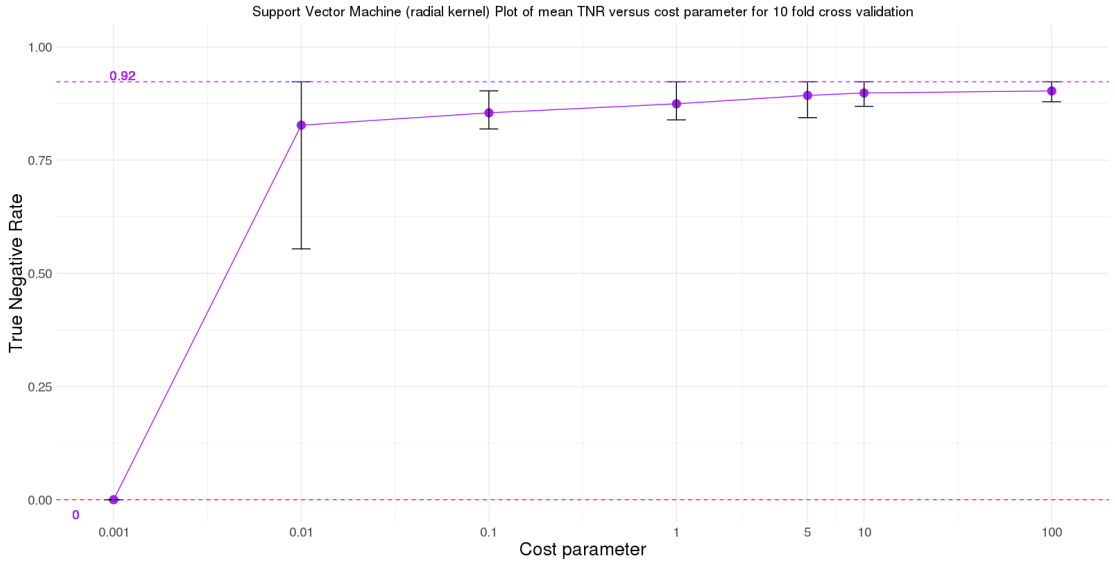

FIGURE A.22: The plot of mean True Negative Rate for Support Vector Machine with a radial kernel within a 10 fold cross-validation for different cost parameters.

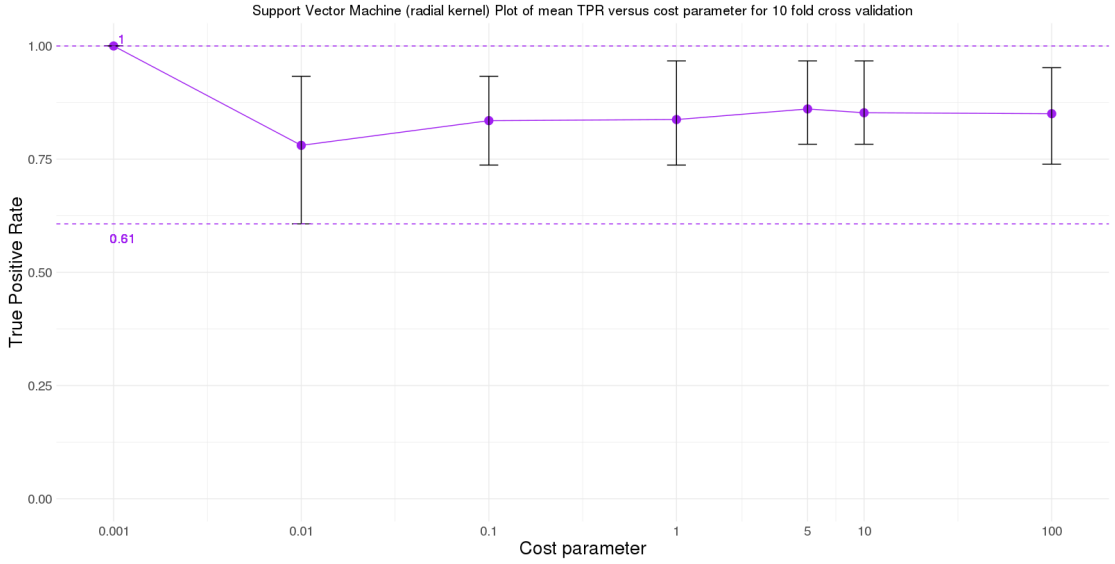

FIGURE A.23: The plot of mean True Positive Rate for Support Vector Machine with a radial kernel within a 10 fold cross-validation for different cost parameters.

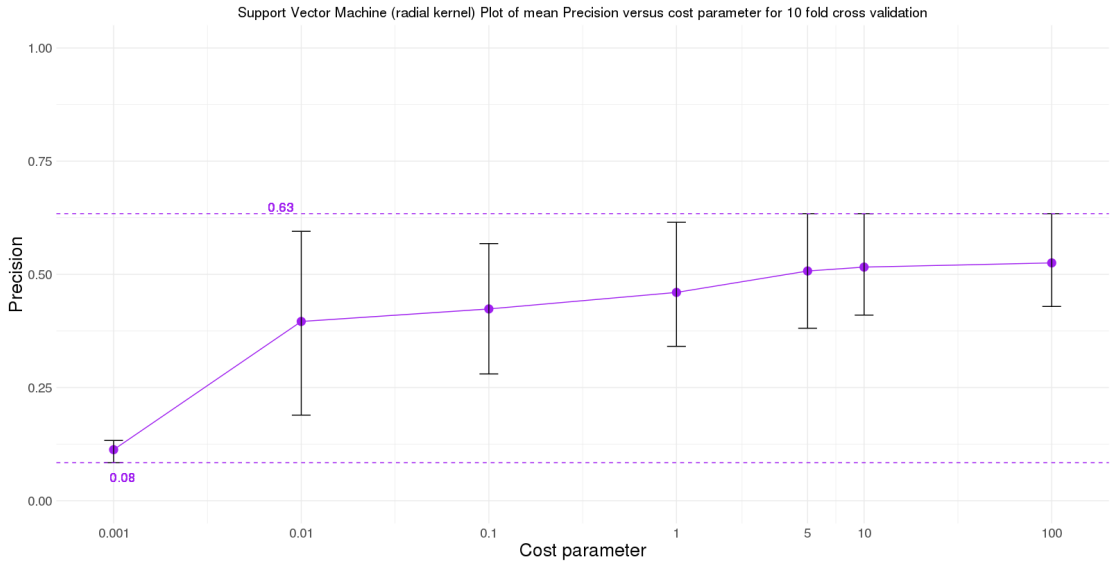

FIGURE A.24: The plot of mean Precision for Support Vector Machine with a radial kernel within a 10 fold cross-validation for different cost parameters.

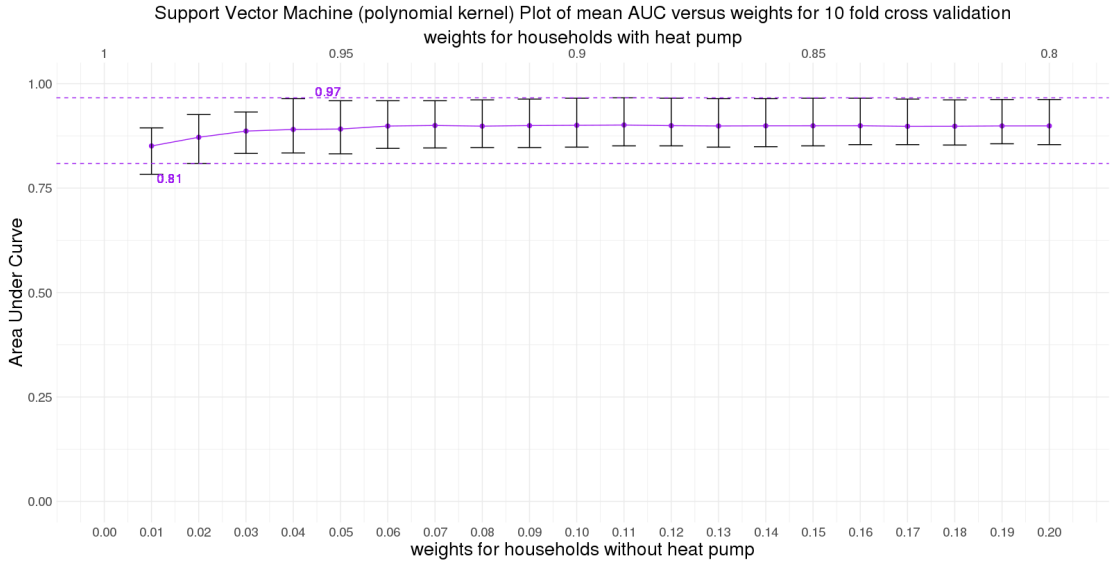

FIGURE A.25: The plot of mean Area Under Receiver Operating Characteristic for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different class weight parameters.

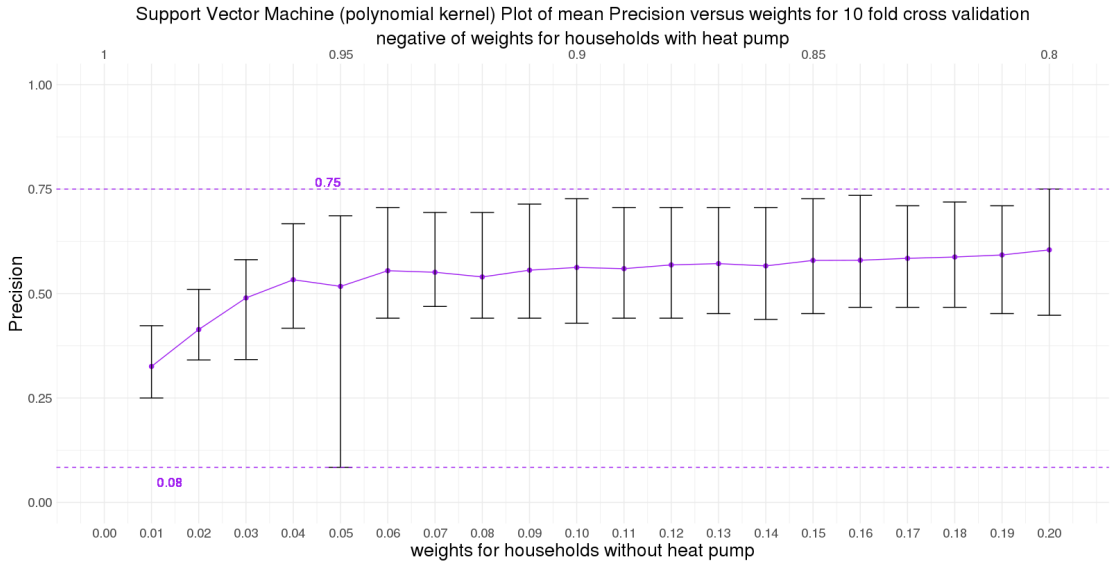

FIGURE A.26: The plot of mean Precision for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different class weight parameters.

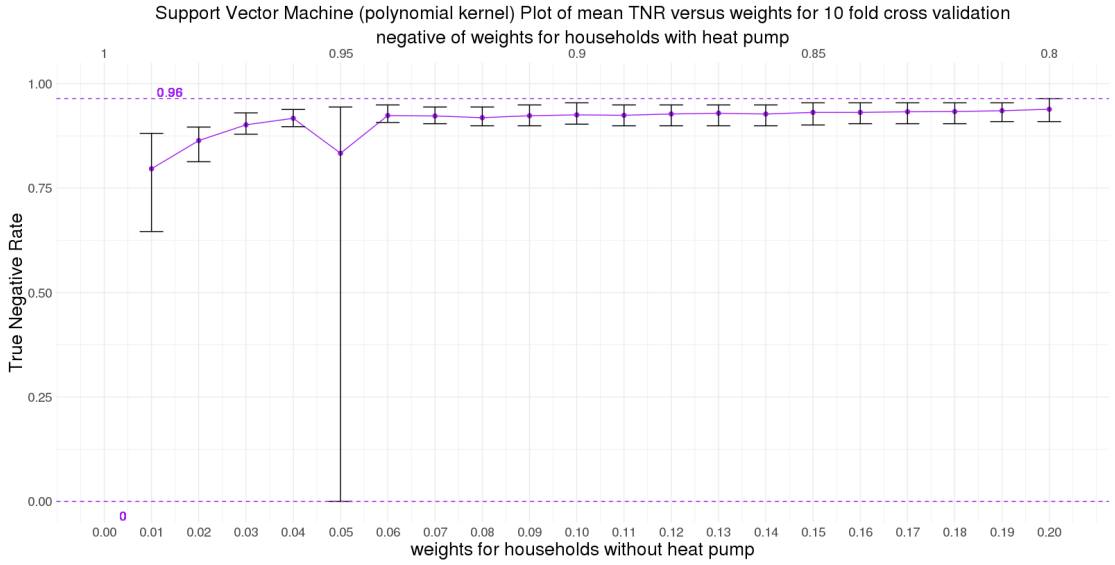

FIGURE A.27: The plot of mean True Negative Rate for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different class weight parameters.

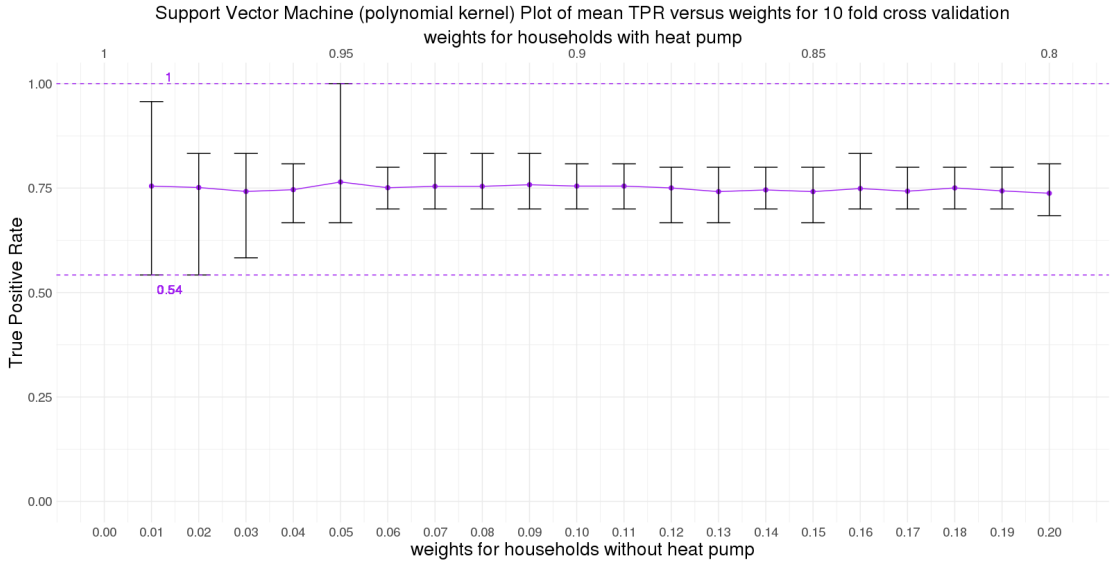

FIGURE A.28: The plot of mean True Positive Rate for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different class weight parameters.

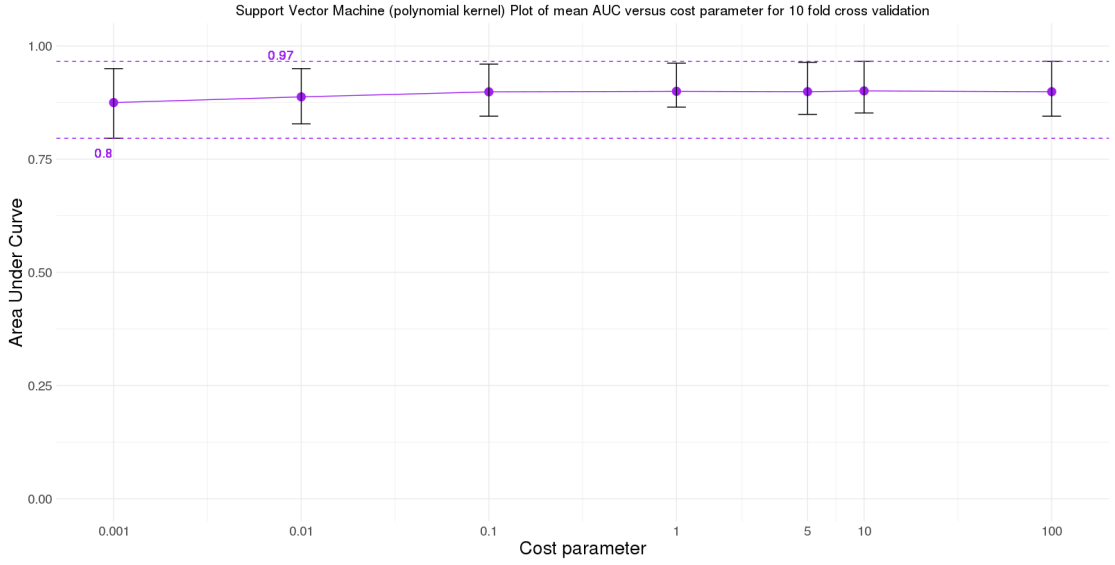

FIGURE A.29: The plot of mean Area Under Receiver Operating Characteristic for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different cost parameters.

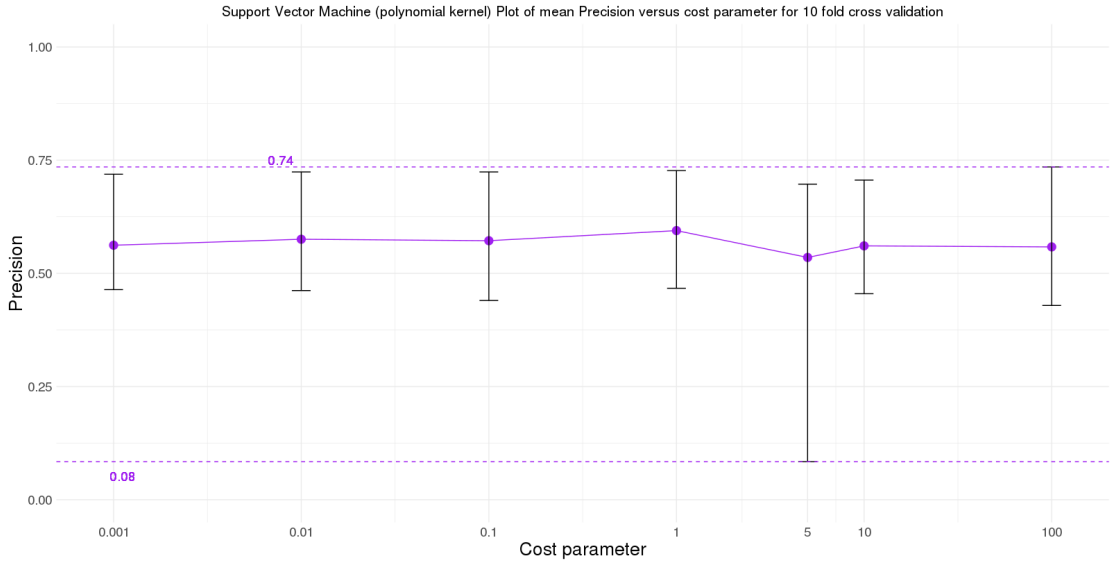

FIGURE A.30: The plot of mean Precision for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different cost parameters.

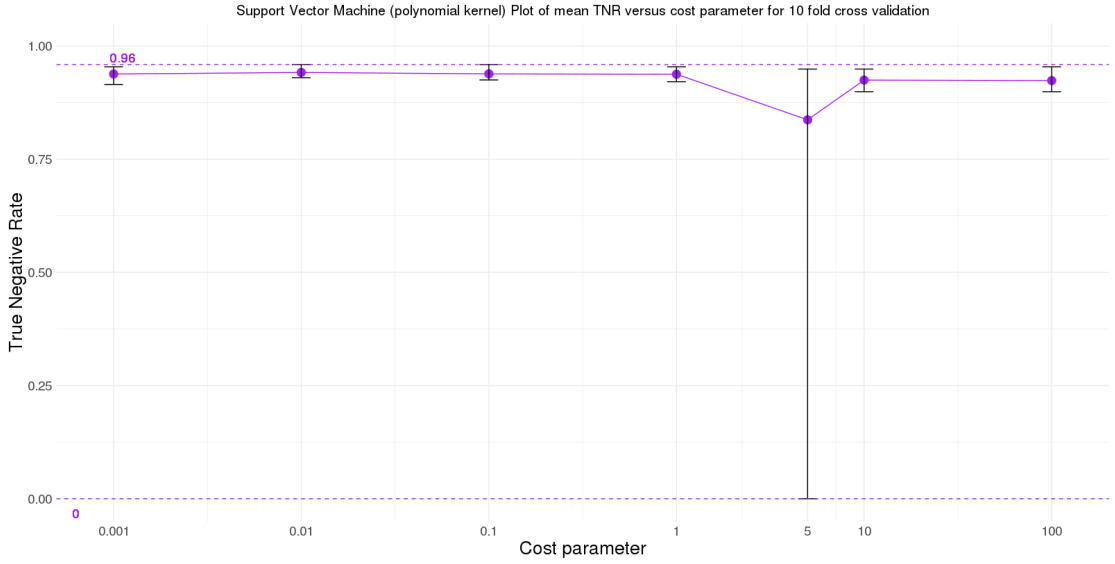

FIGURE A.31: The plot of mean True Negative Rate for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different cost parameters.

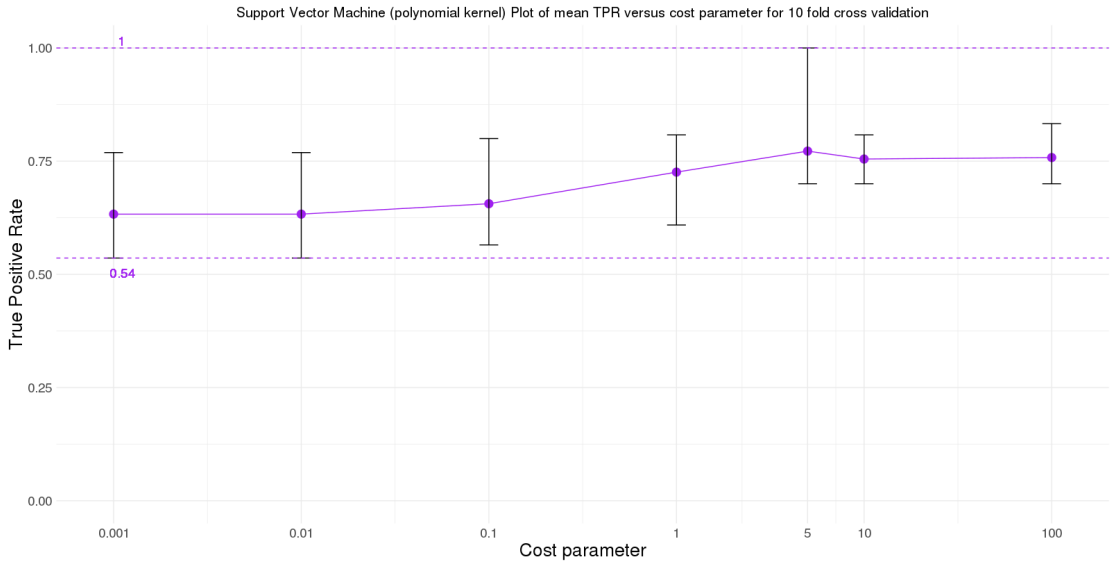

FIGURE A.32: The plot of mean True Positive Rate for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different cost parameters.

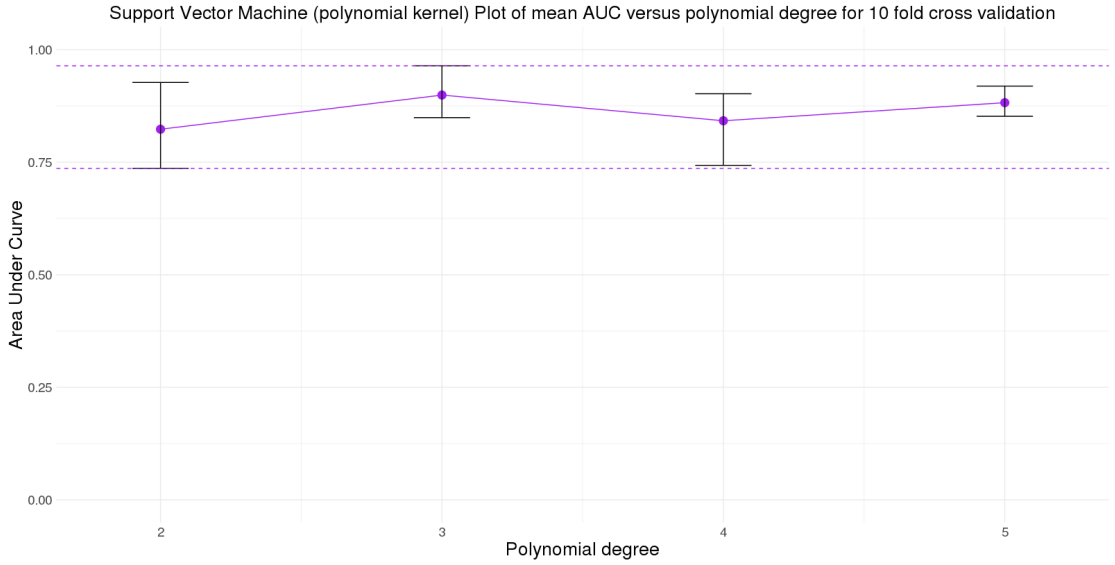

FIGURE A.33: The plot of mean Area Under Receiver Operating Characteristic for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different degrees.

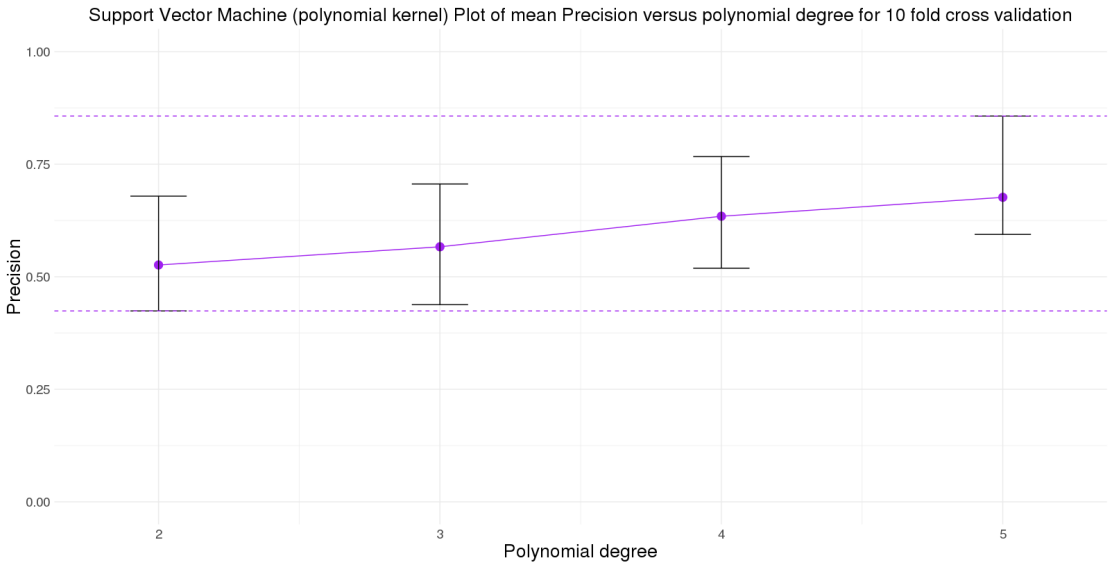

FIGURE A.34: The plot of mean Precision for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different degrees.

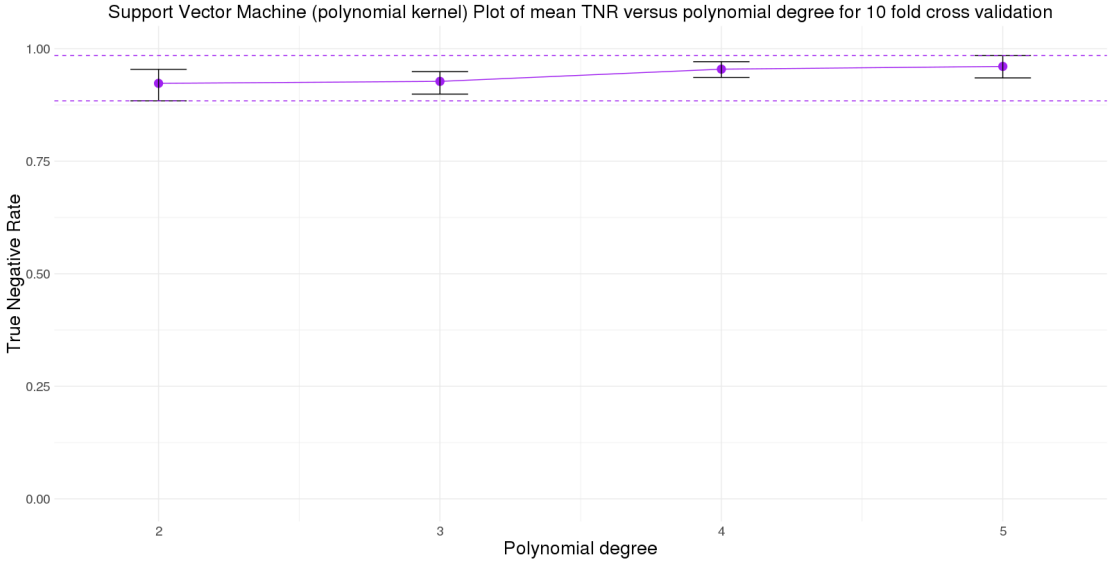

FIGURE A.35: The plot of mean True Negative Rate for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different degrees.

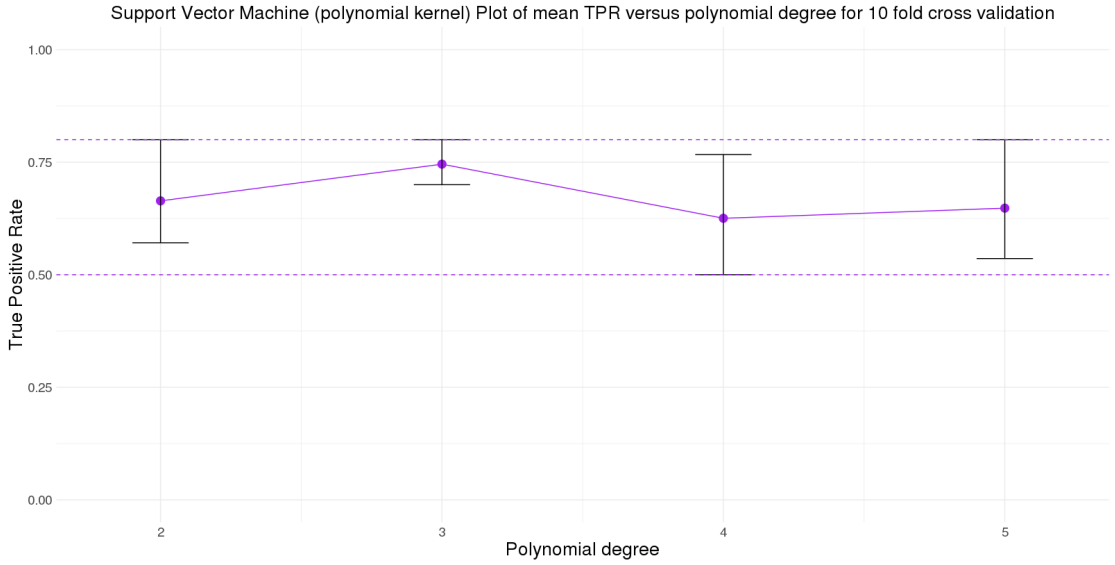

FIGURE A.36: The plot of mean True Positive Rate for Support Vector Machine with a polynomial kernel within a 10 fold cross-validation for different degrees.

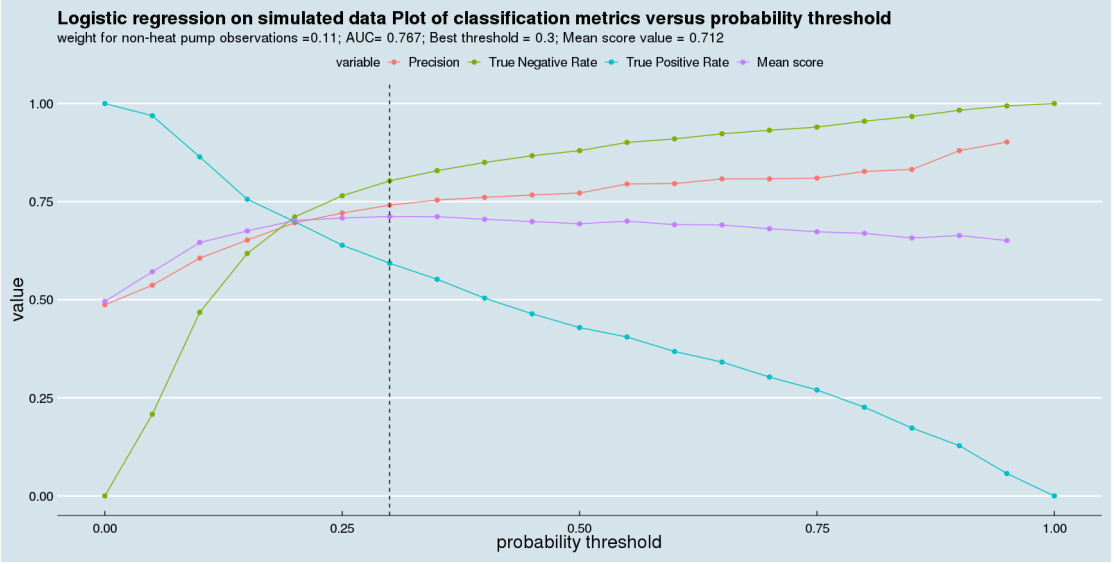

FIGURE A.37: Plot of mean evaluation metrics within 10-fold cross-validation for Logistic Regression with optimal hyperparameters versus probability threshold

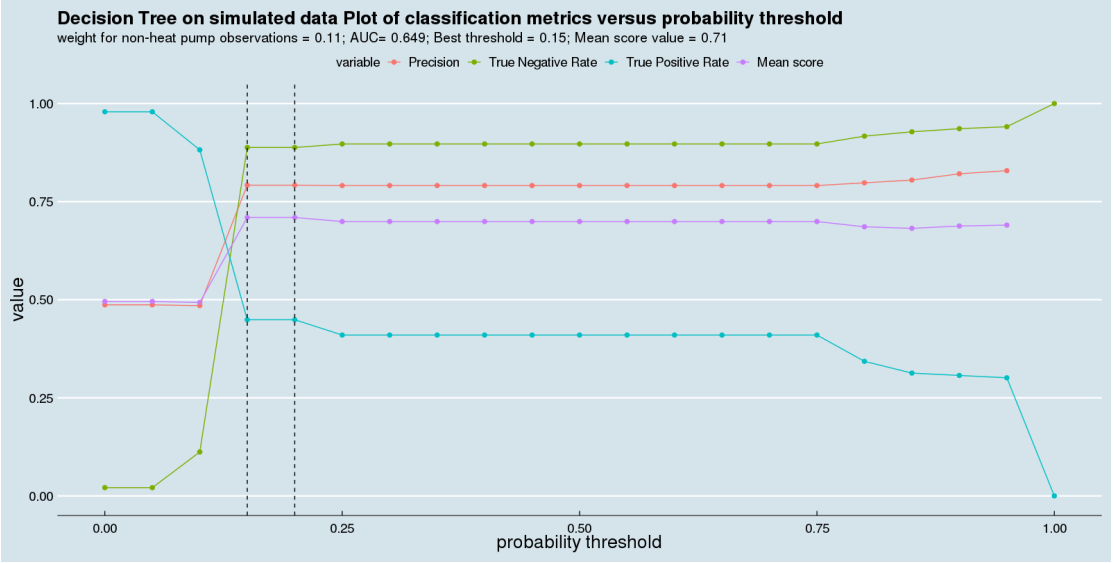

FIGURE A.38: Plot of mean of evaluation metrics within 10-fold cross-validation for Decision Tree with optimal hyperparameters versus probability threshold

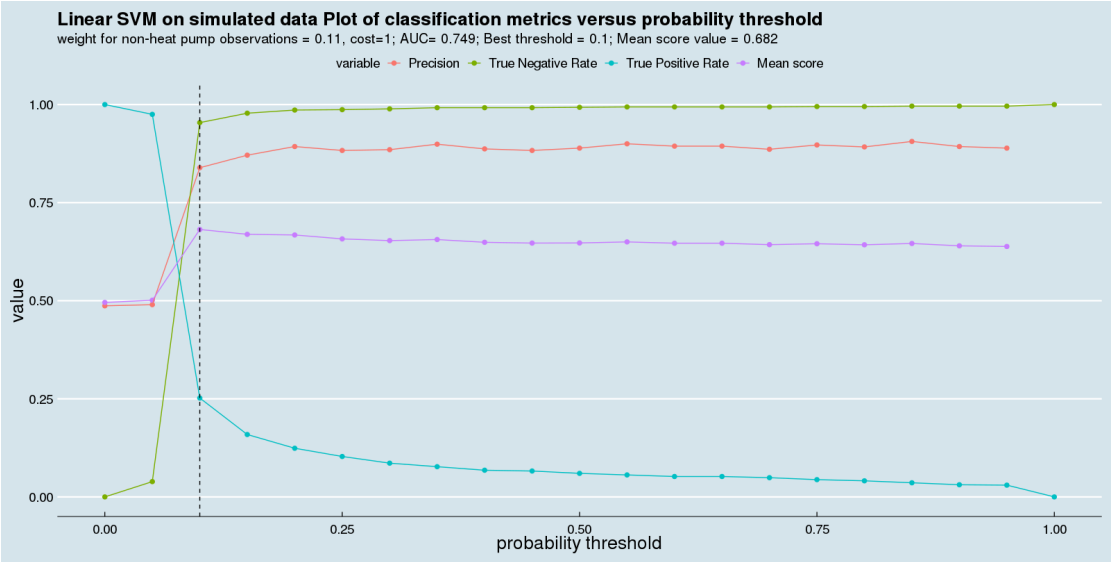

FIGURE A.39: Plot of mean evaluation metrics within 10-fold cross-validation for Linear Support Vector Machine with optimal hyperparameters versus probability threshold

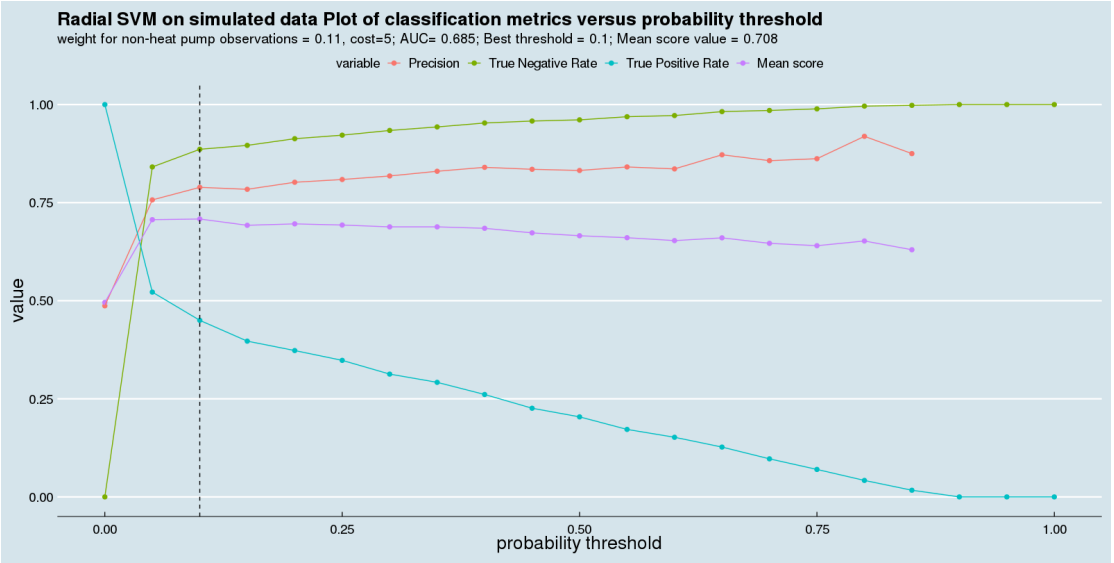

FIGURE A.40: Plot of mean of evaluation metrics within 10-fold cross-validation for Radial Support Vector Machine with optimal hyperparameters versus probability threshold

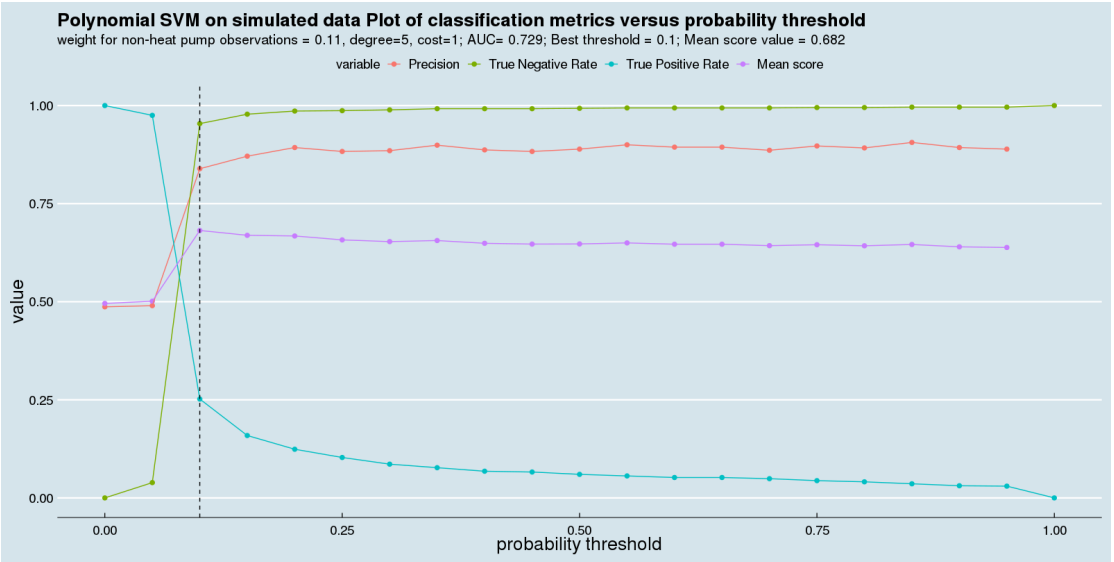

FIGURE A.41: Plot of mean of evaluation metrics within 10-fold cross-validation for Polynomial Support Vector Machine with optimal hyperparameters versus probability threshold

Bibliography

[1] Nur Farahin, Pauzi Abdullah, and Mohammad Yusri. A review disaggregation method in Non-intrusive Appliance Load Monitoring. *Renewable and Sustainable Energy Reviews*, 66:163–173, 2016. ISSN 1364-0321. doi: 10.1016/j.rser.2016.07.009. URL http://dx.doi.org/10.1016/j.rser.2016.07.009.

[2] Akshat Rathi. A 19th-century solution to heat homes is helping the world cut emissions, Sep 2018. URL https://qz.com/1331709/global-sales-of-emissions-reducing-heat-pumps-are-soaring/.

[3] United Nations Convention on Climate Change. Paris Agreement. Technical report, 2015.

[4] Eline van den Ende. A revolution: the netherlands kisses gas goodbye, Jun 2017. URL https://energypost.eu/a-revolution-the-netherlands-kisses-gas-goodbye-but-will-it-help-the-climate/.

[5] Alejandro Navarro, Luis F. Ochoa, and Pierluigi Mancarella. Learning from residential load data: Impacts on LV network planning and operation. *Proceedings of the 2012 6th IEEE/PES Transmission and Distribution: Latin America Conference and Exposition, T and D-LA 2012*, pages 2011–2014, 2012. doi: 10.1109/TDC-LA.2012.6319101.

[6] Jenny Love, Andrew Z.P. Smith, Stephen Watson, Eleni Oikonomou, Alex Summerfield, Colin Gleeson, Phillip Biddulph, Lai Fong Chiu, Jez Wingfield, Chris Martin, Andy Stone, and Robert Lowe. The addition of heat pump electricity load profiles to GB electricity demand: Evidence from a heat pump field trial. *Applied Energy*, 2017. ISSN 03062619. doi: 10.1016/j.apenergy.2017.07.026.

[7] Alejandro Navarro-Espinosa and Pierluigi Mancarella. Probabilistic modeling and assessment of the impact of electric heat pumps on low voltage distribution networks. *Applied Energy*, 127:249–266, aug 2014. ISSN 0306-2619. doi: 10.1016/J.APENERGY.2014.04.026. URL https://www.sciencedirect.com/science/article/pii/S030626191400378X.

[8] Christina Protopapadaki and Dirk Saelens. Heat pump and PV impact on residential low-voltage distribution grids as a function of building and district properties. *Applied Energy*, 192:268–281, apr 2017. ISSN 0306-2619. doi: 10.1016/J.APENERGY.2016.11.103. URL https://www.sciencedirect.com/science/article/pii/S0306261916317329.

[9] Kovvali Manasa Rao, Durga Ravichandran, and Kavi Mahesh. Non-Intrusive Load Monitoring and Analytics for Device Prediction. I, 2016.

[10] Khaled Chahine. Electric Load Disaggregation in Smart Metering Using a Novel Feature Extraction Method and Supervised Classification Electric Load Disaggregation in Smart Metering Using a Novel Feature Extraction Method and Supervised Classification $. (May 2014), 2011. doi: 10.1016/j.egypro.2011.05.072.

[11] Hugo Gonçalves, Adrian Ocneanu, and Range Hood Fan. Unsupervised disaggregation of appliances using aggregated consumption data.

[12] Ereola Johnson Aladesanmi. Non-Intrusive Load Monitoring and Identification for Energy Management Systems Using Computational Intelligence Approach. 2015.

[13] Yi Wang, Qixin Chen, Tao Hong, and Chongqing Kang. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges, 2018. ISSN 19493053.

[14] Usama Fayyad, Gregory Piatetsky-shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[15] Colin Shearer. The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 2000.

[16] Shiyin Zhong and Kwa Sur Tam. Hierarchical Classification of Load Profiles Based on Their Characteristic Attributes in Frequency Domain. *IEEE Transactions on Power Systems*, 2015. ISSN 08858950. doi: 10.1109/TPWRS.2014.2362492.

[17] Amir Kavousian, Ram Rajagopal, and Martin Fischer. Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, 2013. ISSN 03605442. doi: 10.1016/j.energy.2013.03.086.

[18] Ramon Granell, Colin J. Axon, and David C.H. Wallom. Clustering disaggregated load profiles using a Dirichlet process mixture model. *Energy Conversion and Management*, 2015. ISSN 01968904. doi: 10.1016/j.enconman.2014.12.080.

[19] Antonio Ridi, Christophe Gisler, and Jean Hennebert. A survey on intrusive load monitoring for appliance recognition. In *Proceedings - International Conference on Pattern Recognition*, 2014. ISBN 9781479952083. doi: 10.1109/ICPR.2014.636.

[20] Anthony Schoofs, Antonio G Ruzzelli, and Gregory M. P. O'Hare. Appliance activity monitoring using wireless sensors. *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks - IPSN '10*, 2010. doi: 10.1145/1791212.1791292.

[21] A. Schoofs, A. Guerrieri, D. T. Delaney, G. M.P. O'Hare, and A. G. Ruzzelli. ANNOT: Automated electricity data annotation using wireless sensor networks. In *SECON 2010 - 2010 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, 2010. ISBN 9781424471515. doi: 10.1109/SECON.2010.5508248.

[22] Kaustav Basu, Vincent Debusschere, and Seddik Bacha. Appliance usage prediction using a time series based classification approach. In *IECON Proceedings (Industrial Electronics Conference)*, 2012. ISBN 9781467324212. doi: 10.1109/IECON.2012.6388597.

[23] Xiaofan Jiang Xiaofan Jiang, S. Dawson-Haggerty, Prabal Dutta, and David Culler. Design and implementation of a high-fidelity AC metering network. *2009 International Conference on Information Processing in Sensor Networks*, 2009. ISSN 1076-0342. doi: 10.1061/(ASCE)1076-0342(2008)14:1(89).

[24] J Zico Kolter and Matthew J Johnson. REDD : A Public Data Set for Energy Disaggregation Research. In *Proceedings of the 1st KDD Workshop on Data Mining Applications in Sustainability (SustKDD*, 2011. ISBN 9781450308403.

[25] J Zico Kolter, Siddharth Batra, and Andrew Y. Ng. Energy disaggregation via discriminative sparse coding. In *NIPS*, 2010. ISBN 9781617823800. doi: 10.1080/13528165.2012.712256.

[26] Sidhant Gupta, Matthew S. Reynolds, and Shwetak N. Patel. Electrisense: single-point sensing using emi for electrical event detection and classification in the home. In *UbiComp*, 2010.

[27] Shwetak N. Patel, Thomas Robertson, Julie A. Kientz, Matthew S. Reynolds, and Gregory D. Abowd. At the flick of a switch: Detecting and classifying unique electrical events on the residential power line (nominated for the best paper award). In *UbiComp*, 2007.

[28] Valeria Amenta and Giuseppe Marco Tina. Load demand disaggregation based on simple load signature and user's feedback. In *Energy Procedia*, 2015. ISBN 978-0-7918-4394-9. doi: 10.1016/j.egypro.2015.12.213.

[29] Hyungsul Kim, Manish Marwah, Martin Arlitt, Geoff Lyon, and Jiawei Han. Unsupervised Disaggregation of Low Frequency Power Measurements. In *Proceedings of the 2011 SIAM International Conference on Data Mining*. 2011. ISBN 9780898715453. doi: 10.1137/1.9781611972818.64.

[30] Hongliang Fei, Younghun Kim, Sambit Sahu, Milind Naphade, and Sanjay K Mamidipallis. Heat Pump Detection from Coarse Grained Smart Meter Data with Positive and Unlabeled Learning. pages 1330–1338, 2013.

[31] George W. Hart. Nonintrusive Appliance Load Monitoring. *Proceedings of the IEEE*, 1992. ISSN 15582256. doi: 10.1109/5.192069.

[32] Commission for regulation of utilities water and energy - cru ireland. URL https://www.cru.ie/.

[33] Thomas Weibel. Umasstracerepository. URL http://traces.cs.umass.edu/index.php/Smart/Smart.

[34] Smartmeter energy consumption data in london households. URL https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households.

[35] Runming Yao and Koen Steemers. A method of formulating energy load profile for domestic buildings in the UK. *Energy and Buildings*, 2005. ISSN 03787788. doi: 10.1016/j.enbuild.2004.09.007.

[36] A. Capasso, W. Grattieri, R. Lamedica, and A. Prudenzi. A bottom-up approach to residential load modeling. *IEEE Transactions on Power Systems*, 9(2):957–964, May 1994. ISSN 0885-8950. doi: 10.1109/59.317650.

[37] James Gareth, Witten Daniela, Hastie Trevor, and Tibshirani Rober. *An Introduction to Statistical Learning with Applications in R*. 2000. ISBN 978-1-4614-7137-0. doi: 10.1007/978-1-4614-7138-7.

[38] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Mit Press, 2012. ISBN 9780262018258. URL http://www.jstor.org/stable/j.ctt5hhcw1.

[39] Applying decision trees to online learning data. URL http://scalar.usc.edu/works/c2c-digital-magazine-fall-2017--winter-2018/applying-decision-trees-to-online-learning-data.

[40] J. M. Banda, R. A. Angryk, and P. C.H. Martens. Steps Toward a Large-Scale Solar Image Data Analysis to Differentiate Solar Phenomena. *Solar Physics*, 2013. ISSN 00380938. doi: 10.1007/s11207-013-0304-x.

[41] R.; Department of Energy Lowe and Climate Change. Renewable heat premium payment scheme: Heat pump monitoring: Cleaned data, 2013-2015. *UK Data Service*, 2017. doi: 10.5255/UKDA-SN-8151-1.

[42] Smartmeter energy consumption data in london households. URL https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households.

[43] KNMI - Hourly weather data in the Netherlands. URL https://www.knmi.nl/nederland-nu/klimatologie/uurgegevens.

[44] Appleby parva: Centre of the country! URL https://web.archive.org/web/20071123015253/http://www.applebymagna.org.uk/population_centre.htm.